Vol. 00, No. 0, Xxxxx 0000, pp. 000–000
ISSN 0000-0000 | EISSN 0000-0000 | 00 | 0000 | 0001

INFORMS
DOI 10.1287/xxxx.0000.0000

Hospital Inpatient Operations: Mathematical Models and Managerial Insights

Pengyi Shi
H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30332, pengyishi@gatech.edu

Mabel C. Chou
Department of Decision Sciences, NUS Business School, National University of Singapore, mabelchou@nus.edu.sg

J. G. Dai
School of Operations Research and Information Engineering, Cornell University, Ithaca, NY 14853; on leave from H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30332, jim.dai@cornell.edu

Ding Ding
School of International Trade and Economics, University of International Business & Economics, Beijing, dingd.cn@gmail.com

Joe Sim
NUS Yong Loo Lin School of Medicine and NUS Business School, National University of Singapore, and National University Hospital, joe_sim@nuhs.edu.sg

One key factor contributing to the emergency department (ED) overcrowding is prolonged waiting time for admission to inpatient wards, also known as ED boarding time. To gain insights into the inpatient flow management to reduce this waiting time, we study the operations in the inpatient department of a Singaporean hospital. We focus on understanding the effect of an early discharge policy, implemented in late 2009, on the fraction of patients who have to wait in ED for six hours or longer to be admitted. Based on a comprehensive empirical analysis of the inpatient department [40], we propose a novel stochastic network model with the following characteristics to model the inpatient operations: (1) A patient's service time is endogenous, depending on her admission and discharge times, and her length of stay. As a consequence, the service times are not independent, identically distributed. (2) Pre- and post-allocation delays for each patient's bed-request, even if a bed is available at the time of request, allow modeling secondary bottlenecks such as temporary nurse shortage. (3) Patients waiting for a bed can be overflowed to a non-primary ward when the overflow trigger time reaches a certain threshold, where the threshold is time-dependent.

We show, via simulation studies, that our model is able to approximately replicate the hourly performances (e.g., waiting time) of the inpatient operations at this hospital. The model allows one to evaluate the impact of operational policies on waiting times and overflow proportions. In particular, our model predicts that implementing a hypothetical *Period 3* policy can eliminate the excessive waiting for those patients who request beds in mornings. The policy constitutes the following components: a discharge distribution with the first discharge peak between 8 and 9am and 26% of patients discharge before noon, and stable-mean allocation delays throughout the day. Although Period 3 policy is not completely practical, it can serve as a goal for hospital managers to aim at.

Key words: inpatient flow management, early discharge, waiting time, stochastic network model, ED overcrowding

1. Introduction

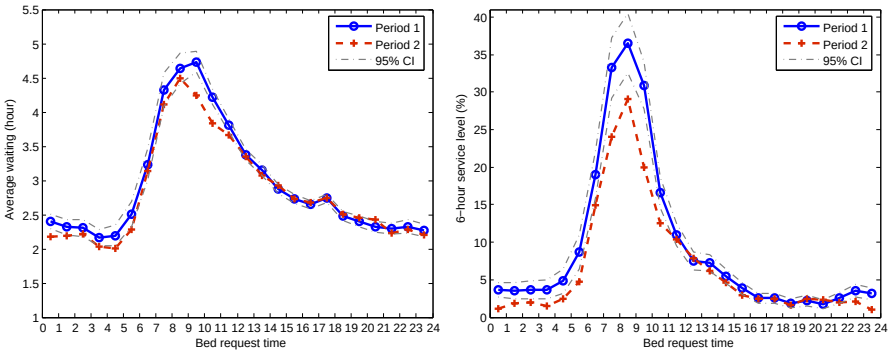
National University Hospital (NUH) is one of the major public hospitals in Singapore. It operates a busy emergency department (ED) and a large inpatient department that has about 1000 beds as of January 1, 2011. NUH, along with all other public hospitals in Singapore, provides weekly report on *waiting time for admission to ward* to the Ministry of Health (MOH); see MOH website for the latest daily waiting time statistics from Singaporean public hospitals [41]. According to the MOH website, the waiting time for admission to ward is computed from the time "Decision by doctor to admit patient" (i.e., time of bed-request after treatment in ED) to the "Time patient exits ED" to go to an inpatient ward. It is also known as the *ED boarding time* in US health systems; see, for example, the definition in Table 1 of [46]. In this paper, we adopt a slight different definition when reporting the empirical waiting time statistics. We use the admission time to wards as the end point of waiting. Thus, our reported waiting time is a slight overestimation of the one in [41], since the gap between patient exiting ED and admission to ward is about 18 minutes on average.

Waiting time at NUH

From January 1, 2008 to June 30, 2009, labeled as Period 1 in this paper, the average waiting time at NUH is 2.82 hours (169 minutes). Unless otherwise specified, the waiting time is always calculated for ED-GW patients, who have been treated in the emergency department and then admitted into a general ward (GW). (General wards are defined precisely in Sections 2.1 and 2.4 of [40]; they exclude many specialized wards including the intensive care units (ICU). At NUH, around 20% of patients visiting ED become ED-GW patients.)

The 2.82 hours average waiting time does not seem to be long given the average of 2.27 hours needed by the bed management unit (BMU) to allocate an inpatient bed plus the time to transport a patient to GW. (See Section 3.2 for a discussion on allocation delays.) This level of complacency immediately evaporates if we examine the waiting times of patients requesting beds in mornings. The blue curve in Figure 1a shows that the average waiting time is more than 4 hours long for patients who request a bed between 7 and 11am. Moreover, the blue curve in Figure 1b shows that for those patients who request a bed between 7 and 10am, more than 30% of them have to wait 6 hours or longer. In Figure 1, each dot over an hour represents the statistics compiled from patients who made bed-requests during that hour, and the 6-hour service level is defined as the fraction of patients who have to wait 6 hours or longer.

While no patient likes waiting, 6 hours or more is extremely undesirable for the following reasons: (1) admitted patients may be forced to wait in areas affording little or no privacy [35], and such a long wait can be frustrating; (2) providers may miss the best timing for treatment; (3) patients complain to hospital management and overseers, resulting in possible adverse publicity (e.g., see local newspapers articles on prolonged waiting time [44]). Thus, it is important for hospitals to eliminate the excessive amount of waiting and achieve *time-stable* performances, even though the bed-request rates of ED-GW patients are usually time-varying. See, for example, the green curve in Figure 8 in Section 3.5.1, the bed-request patterns in [37], and Figure 6 of [1].



(a) Average waiting times (b) 6-hour service level

Figure 1 **Hourly waiting times statistics for ED-GW patients; Period 1: January 1, 2008 to June 30, 2009; Period 2: January 1, 2010 to December 31, 2010. Each dot represents the average waiting time or 6-hour service level for patients requesting beds in that hour. For example, the dot between 7 and 8 represents the value of the hourly statistics between 7am and 8am. The 95% confidence intervals are plotted for Period 1 curves.**

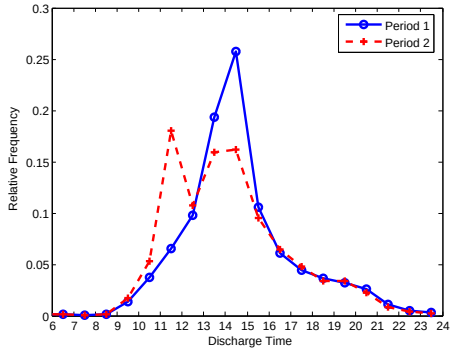


Figure 2 **Discharge time distributions in Periods 1 and 2. The values in the first 6 hours are nearly zero and are not displayed. Readers are referred to Section 3.1 of [40] for the corresponding numerical values.**

Early discharge campaign and discharge distribution

The hourly waiting time statistics depend on the discharge pattern of patients. The blue curve in Figure 2 plots the discharge distribution of patients from general wards in Period 1; for each hour, the corresponding dot represents the fraction of all patients who are discharged during that hour in the period. Clearly, the peak discharge hour is between 2pm and 3pm, and therefore many admissions have to occur after 3pm. In other words, if there is no bed immediately available for a morning bed-request, the patient is likely to wait until afternoon to be admitted since only 12.7% patients are discharged before noon. This discharge pattern was believed by NUH, and many other hospitals alike, to have contributed to the excessively long waiting for patients who request beds in the morning, between 7am and 11am.

In July 2009, NUH launched an early discharge campaign to reduce the waiting time for ED-GW patients and to alleviate the pressure of overflowing patients into non-primary wards (see more discussions on the latter in Section 2.3). After six months' implementation, a new discharge pattern has emerged in Period 2: January 1, 2010 to December 31, 2010. The red curve in Figure 2

displays the new discharge distribution. A morning discharge peak arises, occurring between 11am and noon; 26% of the patients now are discharged before noon in Period 2, more than double of the proportion in Period 1.

Notably, the Period 1 discharge pattern is not unique at NUH; see, for example, Figure 6 of [1]. Discharges in hospitals from many countries are also clustered in the afternoons, while bed-requests from ED-GW patients patients tend to be spread more evenly over the day. Thus, early discharge policies are widely recommended. See references and more discussions in Section 1.2. Intuitively, early discharge from wards should help reduce the waiting time for admission to ward and these recommendations are based on this intuition. Is this intuition right? The red curve in Figure 1a plots the average waiting time at NUH in Period 2, and the red curve in Figure 1b plots the corresponding 6-hour service level. From the red curves, we observe that (a) some improvement in waiting time statistics has been achieved in Period 2, and (b) little progress has been made in achieving a time-stable performance despite the early discharge campaign.

The plots in Figure 1 raise two issues. First, it is unclear whether the improvements in Period 2 result from the NUH's early discharge campaign. As in many hospitals, the operating environment is continuously changing at NUH. Bed capacity is increasing in response to the rising number of patients seeking treatment. In Period 2, the overall bed occupancy rate (BOR) has reduced by 2.7%; see the empirical analysis in Section 3.3 of [40]. Therefore, it is impossible to evaluate the impact of the early discharge policy through pure empirical analysis. Second, one wonders if there is any discharge policy, perhaps combining with other operational improvements, that can achieve a time-stable performance, especially given the time-varying bed-request pattern from ED-GW patients. Unfortunately, it is prohibitively expensive for hospitals to experiment with various options in a real operational environment to identify such policies. Therefore, a high-fidelity model is needed to resolve these concerns.

Contributions

This paper makes two major contributions to the modeling and practice of inpatient flow management. First, it develops a new stochastic network model that strikes the right balance between tractability and relevance. Second, the model provides a set of useful managerial insights for inpatient operations.

For the first contribution, our stochastic network model is tractable because most of its input distributions and parameters can be obtained from an empirical analysis, which is documented in the online supplement [40]. It is relevant because we are able to approximately replicate many performance measures at both the hospital and the medical specialty levels. In particular, we replicate the hourly waiting time performances. In order for the model to be relevant, several key features must be built into the model. They include pre- and post-allocation delays, *endogenous* service times, and dynamic overflow trigger time. These novel features are now briefly discussed in the next three paragraphs and will be elaborated further in Section 4.

The first feature is that each patient who requests a bed needs to experience a pre-allocation delay before a bed can be allocated, and a post-allocation delay before actual occupying the allocated bed. The pre-allocation delay includes the time that BMU needs to search and negotiate for a bed from a ward and the post-allocation delay includes part of the time needed for the patient to be discharged from ED and the transportation time from ED to the allocated bed. A number of factors influence these allocation delays. For example, the lack of BMU agents, ward physicians, and ward nurses extends the pre-allocation delay; the lack of ED physicians, ED nurses, and porters (escorting patients from ED to wards) extends the post-allocation delay. Empirical analysis in Section 4.1 shows that these allocation delays are time-varying, depending on the hour of a day.

The second feature is that the service time of a patient is endogenous. It is dictated by admission time and two exogenous variables: length of stay (LOS), and discharge time. Here, both admission and discharge time refer to the time of the day; LOS equals the number of nights in the hospital stay. As a consequence, the service times are *not* independent, identically distributed (iid). Section 4.3 will explain an alternative, exogenous service time model that fails to reproduce the discharge distribution of patients and other hourly performance measures.

The third feature is that when the waiting time of a patient exceeds a critical value called the overflow trigger time, she can be overflowed to a non-primary ward; the overflow trigger time can be time- and state-dependent. See Section 4 for a detailed description of the necessity of including all the three features when predicting the time-dependent waiting time statistics. As we mentioned, excessively long waiting time is extremely undesirable and the focus of this paper is to stabilize the waiting time performances. Therefore, it is crucial for our model to be able to replicate the hourly performances, not just the daily ones.

For the second contribution of this paper, we obtain a number of managerial insights through simulation analysis of the stochastic model. First, the early discharge alone, at the level achieved at NUH in Period 2, has little impact on stabilizing the waiting time of ED-GW patients; recall that under this discharge distribution, 26% of patients are discharged before noon with the first discharge peak occurring between 11 am and 12 noon. Second, if the hospital is able to (i) move the first peak in the Period 2 discharge distribution three hours earlier (between 8am and 9am) and still keep only 26% discharge before noon, and (ii) meanwhile achieve time-stable pre- and post-allocation delays, with a constant mean (about 1 hour) for each allocation delay distribution at each bed-request hour, then a time-stable waiting time performance can be attained (see Figure 19). See Section 5 for additional managerial insights generated from our model.

To our best knowledge, this paper is the first to build a stochastic model to analyze the effect of discharge policy and allocation delays. In a recent paper [37], the authors propose a deterministic fluid model to analyze the effect of discharge timing on ED boarding times. One does not expect such kind of model to accurately capture the waiting time statistics because it is well known that average waiting time depends on the variability of interarrival and service times [19]. The authors

find that by shifting the peak inpatient discharge time four hours earlier, ED boarding time (i.e., waiting time for ED-GW patient in our paper) is essentially reduced to zero. We cannot replicate their findings in our model of NUH inpatient department. In addition to being deterministic, their model does not account for allocation delays. (They assume a deterministic two-hour bed-cleaning delay after each discharge, which is different from pre- and post-allocation delays in our model.) Therefore, their model cannot capture the interplay between bed shortage and allocation delays caused by *secondary bottlenecks* due to the imbalance between manpower availability and workload across different hour of the day. Readers are referred to Section 1.2 for more references on discharge policy and patient flow modeling. Regarding secondary bottlenecks, we are not the first one to discover their impact on waiting time for admission to ward. Yankovic and Green [51] point out that nurse availability adds another resource constraint besides the inpatient bed availability. They propose a new two-queue model to determine nurse staffing level in an inpatient ward and demonstrate that admission or discharge *blocking* caused by nurse shortage can have a significant impact on system performances.

Interfaces with other hospital departments

Much literature has studied ED operations (see Section 1.2). There is no doubt that efficiency of the ED operation directly affects waiting experience and the medical safety of a patient. In this study, we intentionally choose not to model ED operations to keep our model tractable. Instead, we choose to model the interface between ED and inpatient department carefully, for example, the time-varying allocation delays. Our proposed model on inpatient flow management in this paper can be used for other studies that need to integrate the ED and inpatient department operations together.

We do not model the ICU-type wards explicitly, because the data requirements to model them would be at another level and are beyond the scope of this paper. Alternatively, we use one admission source, ICU-GW patients, to model the interactions between ICU-type wards and general wards. Sections 3.1 and 3.4 elaborate the details of modeling transfers/re-admissions between ICU-type wards and general wards by patients from this source. Moreover, the main focus of this paper is to stabilize the waiting time performances for ED-GW patients. We have done sensitivity analyses and find that ICU-GW patients have limited impact on these performance measures, partly because the volume of ICU-GW patients is small compared to that of ED-GW patients. Thus, we believe the proposed model is adequate to understand our main focus and to generate correct managerial insights. There are recent studies on ICU wards [8, 30]. Our model can be expanded in future studies to understand the interplay among ED, ICU and general wards.

Apologies

In Armony et al. [1], the authors make the following two apologies in Section 1.1 of their paper: "target audience and space considerations render secondary the role of rigorous statistical analysis", and the "data originates from a single hospital" so some findings may not be universally applicable.

These two apologies apply to our paper as well. For example, in our model we assume that arrival process for each source of patients is non-homogeneous Poisson although this assumption cannot pass rigorous statistical tests for some sources using the empirical data.

1.1. Data set and paper outline

Data set

In the online supplement [40], the authors conduct an empirical study of NUH inpatient operation based on data from January 2008 to December 2010. As mentioned earlier, from July 2009 to December 2009, the early discharge policy was being implemented. The authors exclude this 6-month data, from July 1, 2009 to December 31, 2009, in their analyses of data. Therefore, the data set is separated into two periods. Period 1 is from January 1, 2008 to June 30, 2009, and Period 2 is from January 1, 2010 to December 31, 2010. Period 1 is one and half year long (547 days) and Period 2 is one year long (365 days). When there is no need to separate the data, the authors use a *combined data set* that combines the data from these two periods. When a period is not specified, the combined data set is used. Detailed description of the data sets appears in Section 2.5 of the online supplement [40].

An outline

The remainder of this paper is organized as follows. Related work is summarized in Section 1.2. In Section 2, we give a brief description of the NUH inpatient department. In Section 3, we propose a stochastic network model that captures the basic operation of the NUH inpatient department. A few key features of the stochastic model are elaborated in Section 4. In Section 5, we use our stochastic model to generate a number of important managerial insights for reducing and stabilizing waiting times for admission to wards. The paper concludes in Section 6.

1.2. Literature review

ED overcrowding and early discharge policy

ED overcrowding is a world-wide problem. It compromises patient safety and quality of care; see summaries in two recent review papers [3, 25]. Attention from various areas has been attracted to identify strategies to alleviate ED overcrowding, including areas of emergency medicine, public health, and more recently, operations research [12]. While the past efforts to reduce ED overcrowding have focused on improving activities within the ED, it becomes increasingly clear that outside activities, especially downstream (inpatient) bed availability, have critical impacts on ED crowdedness [2, 27, 31, 38, 49]. Corresponding solutions include changing elective admission schedule to smooth bed occupancy level [23], improving inpatient bed management [27, 38], and balancing inpatient discharges and admissions [48].

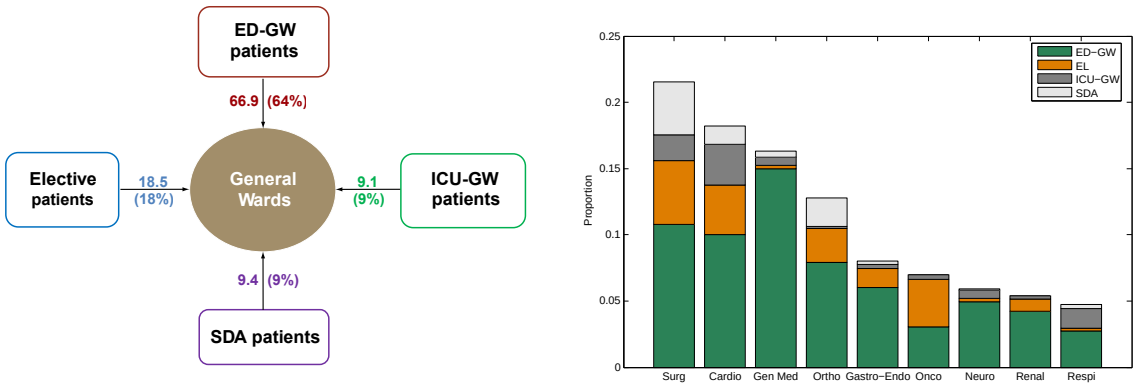
It has been widely perceived that the discharge pattern is one of the main factors leading to prolonged waiting time for admission to wards (i.e., ED boarding time). In many hospitals, discharges are clustered in the afternoon, which causes a temporary mismatch between demand (bed-requests) and supply (available beds) for bed-requests in the morning [11, 37]. One proposed way to eliminate

this mismatch is to use early discharge, i.e. discharging patients in earlier hours of the day [11]. The policy has been recommended by many previous studies [4, 50] and government agencies [11], even though no rigorous studies were conducted to evaluate its impact. Only a few papers, such as [37], have used model to study the relationship between discharge timing and ED boarding time. Other relevant works are mostly empirical studies. For example, [29] classifies admission data from 23 Australian hospitals into five categories based on the relative timing of daily admission and discharge curves. The authors use statistical analysis to show that days with late discharge peaks (more than five hours after the admissions peak) contribute significantly to ED overcrowding. Furthermore, despite that early discharge policy has been recommended for a long time, few hospitals have reported to implement the policy with any success. To our best knowledge, NUH is one of the few hospitals that have successfully implemented the early discharge policy in the *entire* hospital and achieved satisfactory compliance rate. See Section 3.2 of the online supplement [40] for details of NUH's implementation and references on other hospitals which claimed to have implemented or tried to implement the early discharge policy.

Patient flow model and relevant research on call center

Hospital patient flow has been studied extensively in the operations research literature. For example, [1] and [21] conduct detailed studies of patient flow in various departments at an Israeli and a US hospital, respectively. Readers are also referred to many articles cited in these two papers for further references. Armony et al. [1] do not focus on discharge policies, but they empirically study the transfer process flow from ED to GW, which they call internal wards. Discrete-event simulation and queuing theory are two commonly used approaches for modeling and improving patient flow [15, 28, 52]. Comparing to the rich literature on patient flow models of ED, inpatient flow management and the interface between ED and inpatient wards have received less attention; see the same discussion in Section 4 of [1]. Related works on inpatient operations include capacity allocation and flow improvement in specialized hospitals or wards [6, 9, 17, 18], ward nurse staffing [47, 51], bed assignment and overflow [33, 45], and elective admission control and design [23, 24]. Comparing to this line of work, our proposed stochastic network model has many novel components as introduced in the Contributions section of this introduction.

While the daily average or median of the waiting times (for admission to wards) are widely used performance measures when studying ED overcrowding and patient flow model [37, 46], it is not the case for the time-dependent waiting time statistics (see Figure 1). In particular, there has been no research in stabilizing waiting times throughout of the day for admission to wards. The research on call center, however, has extensively studied systems with time-varying performances. For example, Feldman et al. [13] and recent work by Liu and Whitt [32] propose staffing algorithms to achieve time-stable performances. Unlike call center models, our hospital model has extremely long service times with an average of about five days. Within the service time of a typical patient, the arrival pattern has gone through five cycles. Therefore, existing approximation methods developed for call center models are not applicable to our hospital model. Moreover, the servers in our model are inpatient beds. It is not realistic to adjust the number of beds in a short time window.



(a) Admission sources and the daily admission rates (b) Patient distribution based on medical diagnosis

Figure 3 Four admission sources to general wards and nine patient specialties. Daily admission rates and patient distributions are estimated from the combined data.

2. NUH inpatient department : some empirical observations

This section briefly described the operations of the NUH inpatient department. We focus on 19 *general wards*, which is defined precisely in the online supplement [40]. These general wards have a total number of beds ranging from 555 to 638 between January 1, 2008 and December 31, 2010. They exclude a certain number of wards including intensive-care-unit (ICU) wards, isolation wards, high-dependence wards, pediatric wards, and obstetrics and gynaecology (OG) wards. All exclusions are explained in Section 2.4 of [40].

2.1. Admission sources

We classify inpatient admissions to general wards (GWs) into four sources. They are ED-GW, ICU-GW, Elective (EL), and SDA patients. ED-GW patients are those who have completed treatments in the ED and need to be admitted into a general ward. ICU-GW patients are those patients who are initially admitted to ICU-type wards (from either ED or other external resources) and are later transferred to general wards. Most of the Elective (EL) and same-day-admission (SDA) patients come to the hospital to receive surgeries. They are admitted via referrals from clinical physicians, and usually have less urgent medical conditions than ED-GW or ICU-GW patients. The difference between EL and SDA patients is that EL patients are usually admitted in the afternoons *before* the day of surgery, whereas SDA patients first go to the operating room to receive surgery (usually in the morning). *After* the surgery, SDA patients stay temporarily in the SDA ward, typically for a few hours, and then are admitted to a general ward. Therefore, it is expected that an EL patient typically stays in a general ward bed at least one day longer than a SDA patient.

Figure 3a shows the four admission sources and their average daily admission rates which are estimated from the combined data set. In this figure and in all the empirical analysis [40], each patient is only counted once when we calculate the admission rate, even though some patients may be transferred out of and back into general wards after the initial admission. In the model, however, some transfers back into general wards are modeled as a separate stream of pseudo-arrivals (i.e.,

the re-admission class under the ICU-GW source; see Section 3.4 for explanations). Therefore, the daily admission rates to general wards in the model will be higher than the numbers shown in Figure 3a. More discussions on the model input are in Section 3.5. In this paper, patients admitted to general wards from any of the four sources are called *general patients*.

Recall that we define the *waiting time* of an ED-GW patient as the duration between her bed-request time and actual admission time. The average waiting time for all ED-GW patients is 2.82 hours (169 minutes) for Period 1, and 2.77 hours (166 minutes) for Period 2. The overall 6-hour service level is reduced from 6.52% in Period 1 to 5.13% in Period 2. Figures 1a and 1b plot the hourly average waiting time and 6-hour service level in each period. No matter from the overall or hourly waiting time statistics, we only observe modest reduction in Period 2.

For an ICU-GW or a SDA patient, although there is a delay between the bed-request time and the departure time from the ward she currently stays, this waiting time is taken less seriously than that of ED-GW patients. This claim is supported by our empirical observations that the average waiting time is more than 7 hours for ICU-GW patients and about 3.5 hours for SDA patients, both longer than that of ED-GW patients. The major reason could be (a) the ICU-GW and SDA patients have been comfortably receiving care at the current ward, thus this waiting time is not an issue unless there is a bed shortage in ICU-type wards or the SDA ward; (b) MOH does not monitor this performance measure, so the hospital has less incentive to improve it than the waiting time statistics for ED-GW patients.

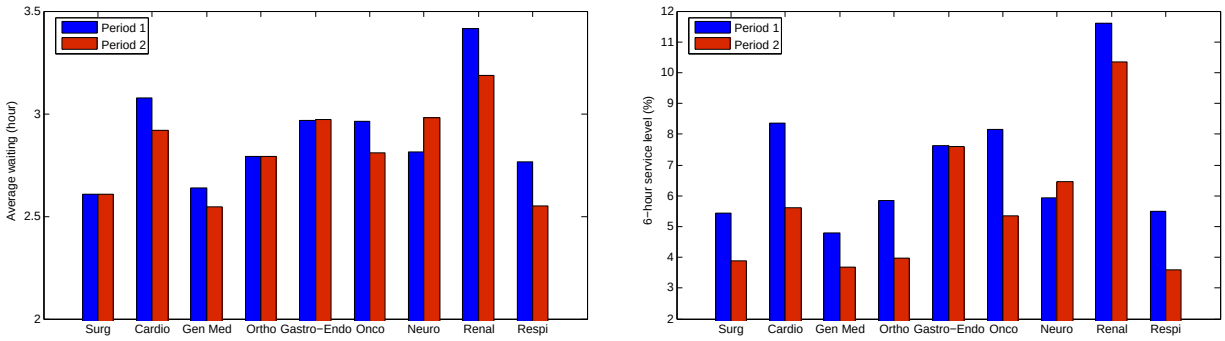
2.2. Medical specialties

General patients are classified by one of nine medical specialities based on diagnosis at time of admission as an inpatient: Surgery, Cardiology, Orthopedic, Oncology, General Medicine, Neurology, Renal Disease, Respiratory, and Gastroenterology-Endocrine. Although Gastroenterology and Endocrine are two different medical specialties, in this paper we group them together and denote as Gastroenterology-Endocrine (Gastro-Endo or Gastro for short). The grouping is based on the fact that patients from these two specialties share the same ward and have similar length of stay (LOS) distributions. See [42] for the same classification. We group Dental, Eye, and ENT patients into Surgery for similar reasons. As explained in Section 2.4 of [40], two other specialties, Obstetrics and Gynaecology (OG) and Paediatrics are excluded from our study.

Figure 3b plots the distribution of general patients among different specialties and admission sources. There is no significant difference in the patient distribution between two periods, so we plot the figure using the combined data. Different specialties show very different admission-source distributions. For example, the majority of General Medicine patients are admitted from ED, while a significant proportion of Surgery patients are EL and SDA patients.

Waiting time statistics for each specialty

Figures 4a and 4b plot the average waiting time and 6-hour service level for ED-GW patients from each specialty in the two periods of study. Renal patients show the longest average waiting time,



(a) Average waiting times (b) 6-hour service level

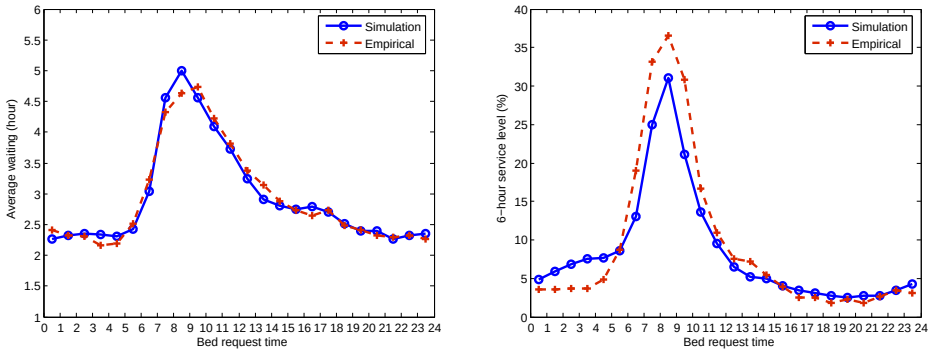
Figure 4 Waiting times statistics for each medical specialty.

and their 6-hour service level is more than 10% in both periods. Surgery, General Medicine and Respiratory patients have better performances on the waiting time statistics than other specialties. Comparing the two periods, the average waiting time remains similar for each specialty, but the 6-hour service levels show a more significant reduction in Period 2 for most specialties, especially for Cardiology and Oncology. Consistent with Figure 1, these observations suggest that the small fraction of patients with long waiting times benefit more in Period 2 than other patients.

2.3. Overflow proportion

In NUH, each general ward is dedicated to a specialty or shared by multiple specialties (see Table 5 in Section 5.1 of the online supplement [40]). We call the ward a *primary ward* for the designated specialty. Usually patients are assigned to their primary wards. However, when an ED-GW patient has waited for several hours in the ED, but no bed from the primary wards is available or expected to be available in the next few hours, NUH may overflow the patient to a non-primary ward as a temporary expedient. Overflow events may also occur among patients admitted from other sources, such as when ICU-type wards need to free up capacity, ICU-GW patients may be overflowed. In this paper and in all the empirical analysis [40], we define the overflow proportion as the number of patients admitted to non-primary wards divided by the total number of admissions. The admissions here include both the initial admission and transfer to general wards, e.g., a transfer from ICU to GW is counted as another admission in addition to the initial admission. Rationales of doing so and more details of the calculations are included in Section 5.2 of [40].

Obviously, there is a trade-off between patient waiting time and overflow proportion. On the one hand, the waiting time can always be reduced by overflowing patients more aggressively since overflow acts as resource pooling. On the other hand, overflow decreases the quality of care delivered to patients and increases hospital operational costs [42]. In NUH, the average overflow proportion among all patients is 26.95% and 24.99% for Periods 1 and 2, respectively. The overflow proportion for all ED-GW patients is 29.91% in Period 1 and 28.54% in Period 2, slightly higher than the values for all patients. The reduction of overflow proportion in Period 2 indicates that the reduced waiting time for ED-GW patients in Period 2 does not result from a more aggressive overflow policy.



(a) Hourly average waiting time (b) Hourly 6-hour service level

Figure 5 Simulation output compares with empirical estimates: hourly waiting time statistics (Period 1).

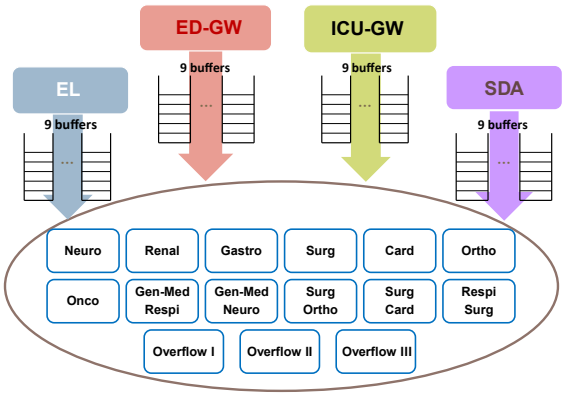


Figure 6 Arrival and server pool configuration in the stochastic model of NUH inpatient department.

Readers are referred to Section 5.2 of the online supplement [40] for discussion on specialty-level and ward-level overflow proportions and additional empirical observations.

3. A stochastic model for the inpatient operations

In this section, we describe our proposed stochastic model. Section 3.1 describes the basic ingredients of the stochastic processing network with multiple-server pools. Section 3.2 describes the details of bed assignments under a specified service policy. Section 3.3 discusses service policies. Under a specified service policy and a specification of model parameters, the stochastic model can be simulated on a computer. The model input is summarized in Section 3.5. In addition, Section 3.4 specifies the details of modeling patient transfer.

Figure 5 shows the simulation estimates and empirical estimates of the hourly waiting time statistics. The simulation estimates are for the baseline scenario, whose definition will be given in Section 5.1. The empirical estimates are obtained from the Period 1 data. The figure shows that our stochastic model can approximately replicate the waiting time performances, even at the hourly resolution.

3.1. A stochastic processing network with multi-server pools

Our proposed stochastic model is a variant of a stochastic processing network that was proposed in Harrison [22] and precisely specified in Dai and Lin [10]. A stochastic processing network processes incoming customers (patients) of various classes. The basic ingredients of a stochastic processing network are servers, buffers, activities, and service policies. Figure 6 depicts a stochastic processing network representation of the NUH inpatient department.

In this paper, general ward beds play the role of servers, and these beds are grouped into $J = 15$ server pools. Each server pool models beds in a general ward or a group of similar wards. We use n_j to denote the number beds in pool j. These n_j beds are assumed to be identical. The 15 server pools serve patients of various types and classes. Here, a *type* is a combination of an admission source and a medical specialty. Since we have four admission sources and nine medical specialities, we have a total of $K = 36$ types. If a server pool serves as a primary pool for patients in at least two specialties, we call it a *shared pool*, and a *dedicated pool* otherwise. See each pool's primary specialties in Figure 6; Section 11.1 of the online supplement [40] explains the three overflow pools in Figure 6. Depending on the admission source of a type, patients within the type may be further segregated into *classes*. Table 1 lists the patient classes within each source for all types. These classes are explained at the end of this section.

In our model, each admission source associates with an arrival process, which is used to model the patient bed-request process. In this paper, we use patient arrival and patient bed-request interchangeably. Each arriving patient (from any of the four sources) is assigned to a specialty (which determines the patient type) with a certain probability that depends both on the source and the arrival hour. Each arriving patient is held in a buffer, waiting to be allocated a bed and later to be admitted into the bed. The waiting patients in these buffers are processed following the service policy described in Section 3.3. Patients within a type are processed in first-in-first-out (FIFO) order.

We assume each type of patients can potentially be assigned to any of the 15 server pools in the model. If a patient is assigned to a primary server pool, we say she is *right-sited*, otherwise, *overflowed*. Adapting the stochastic processing network terminology to the hospital setting, an *activity* is the binding of a server pool serving a particular type of patients. When the server pool is a primary pool for the type, the corresponding activity is said to be a *primary* activity. Clearly, primary activities are more desirable because they avoid patient overflow. However, to reduce waiting time, it is sometimes necessary to activate non-primary activities. A *service policy* dictates which activities to activate at any decision time point. In the hospital setting, a service policy is also known as a *bed assignment policy* that dictates which beds should be assigned to which waiting patients at any decision time point. The decision times have three categories: the bed-request time of a patient, the discharge time of a patient, and the overflow trigger time of a patient. The service policy also dictates the choice of the overflow trigger time for each patient.

A patient can be overflowed only when her waiting time exceeds the overflow trigger time. See Sections 3.3 and 4.2 for details of the service policies used in our model.

Once a patient is admitted into a bed, the patient occupies the bed until discharge. (Discharge in the model corresponds to a final discharge or a transfer-out of a general ward in the hospital; see Section 3.4.) The duration of occupation is called the patient's *service time*. The service time of each patient is random. Unlike traditional queueing models, we propose that service times should not be modeled as exogenous, iid random variables, but rather endogenous variables determined by admission times, LOS distributions, and discharge distributions. Section 4.3 details our proposed service time model.

LOS and discharge distributions depend on patient type (admission source and medical specialty). Moreover, since patients of the same type may still have different LOS or discharge distributions, we further differentiate them by patient *classes*. Empirical study shows that the LOS distribution of ED-GW patients also depends on the admission period (see Section 7.2 of [40]). Here, an admission period is either a before-noon period (from midnight to noon, denoted as AM) or an after-noon period (from noon to midnight, denoted as PM). In addition, some ED-GW and EL patients at NUH transfer from general wards to ICU-type wards after their initial admissions. For such transfer patients, their initial LOS (before transfer to ICU-type wards) are typically shorter than their non-transfer counterparts. Therefore, we divide each EL patient type into *normal* (non-transfer) and *transfer* classes, and divide each type of ED-GW patients into AM-transfer and AM-normal or PM-transfer and PM-normal classes, depending on the admission period.

At NUH, most transfer patients from the ED-GW or EL sources will transfer back to general wards after the stay in ICU-type wards. In other words, such a patient makes the second transfer to a general ward which can be different from the general ward initially occupied. The second transfer of these patients will be modeled as new classes of pseudo-patients from the ICU-GW source. Therefore, we divide each type of the ICU-GW patients into *newly-admitted* and *re-admitted* classes. Newly-admitted patients are those referred in Section 2.1 and in the empirical studies in [40]. Re-admitted patients are those pseudo-patients used to model the second transfer of certain patients. Their LOS distributions are different from those of the newly-admitted ICU-GW patients. See Section 3.4 for the details of modeling patient transfers.

Each patient from a type is assigned to a class following a certain probability that depends on the type and hour of the admission. Patients within each class are homogeneous in terms of LOS and discharge distributions. Table 1 summarizes the classes under each admission source for a given specialty. SDA patients only have one class, and we refer to it as normal for consistency in the table. In the rest of this paper, non-transfer patients refer to those from the normal classes under the ED-GW and EL sources, all patients from the SDA source, and newly-admitted patients from the ICU-GW source. Transfer patients are all other patients.

Source	non-transfer	transfer
ED-GW	normal-AM, normal-PM	transfer-AM, transfer-PM
EL	normal	transfer
ICU-GW	newly-admitted	re-admitted
SDA	normal	

Table 1 Classes under each admission source; same table for each specialty.

3.2. Bed assignment with bed allocation delays

In this section, we spell out the details of our model for bed assignment under a specified service policy when both the pre- and post-allocation delays are present. At NUH, the following four time stamps are registered in its database for each bed-request, and we adopt the same definition in our model. *Bed-request* time is the time when the patient requests a bed. *Bed-request-completion* time is the time when the patient starts to occupy an allocated bed, completing the bed-request. *Allocation-completion* time is the time when a bed is allocated to the patient; the allocated bed is not necessarily available at this time. *Bed-available* time is the time when the allocated bed becomes available (unoccupied). We introduce a new time stamp called *allocation-start time* to model the time when the hospital's bed management unit (BMU) begins the bed allocation process of a patient. This time stamp is *not* recorded at NUH.

When a patient makes a bed-request, our model assumes two allocation modes: *normal allocation* and *forward allocation*. In a normal allocation, the allocation process starts immediately at the bed-request time if a primary bed is available at that time. If no primary bed is available, the patient waits in a buffer for a bed. When a bed becomes available, the allocation process starts. A forward allocation is used only in the second case, i.e., there is no primary bed available at the bed-request time. A forward allocation process starts immediately at the bed-request time. In practice, the allocation process in the second case may start somewhere between the bed-request and bed-available time, and the actual allocation mode is neither normal nor forward as in the model. Due to the lack of accurate time stamps, our model uses these two simplified allocation modes to approximate the reality. We assume that a bed-request at time t has probability $p(t)$ to be a normal allocation and probability $1-p(t)$ to be a forward allocation. In Section 4.1 we will provide more motivation for our model when we discuss how to choose $p(t)$.

The *pre-allocation delay* is the duration between allocation-start time and allocation-completion time, which models the time needed for BMU to search and negotiate for a bed from an appropriate ward. The *post-allocation delay* is the duration between allocation-completion time or bed-available time, whichever is later, and bed-request-completion time. The post-allocation is used to model the delay in discharging patients from ED, and the delay in transporting them from ED or other non-general wards to a general ward. Figures 7 illustrates the relationship of pre- and post-allocation delays with various time stamps.

In our model, we define the waiting time of a patient as the duration between her bed-request (arrival) time and bed-request-completion (admission) time, consistent with what we used to report

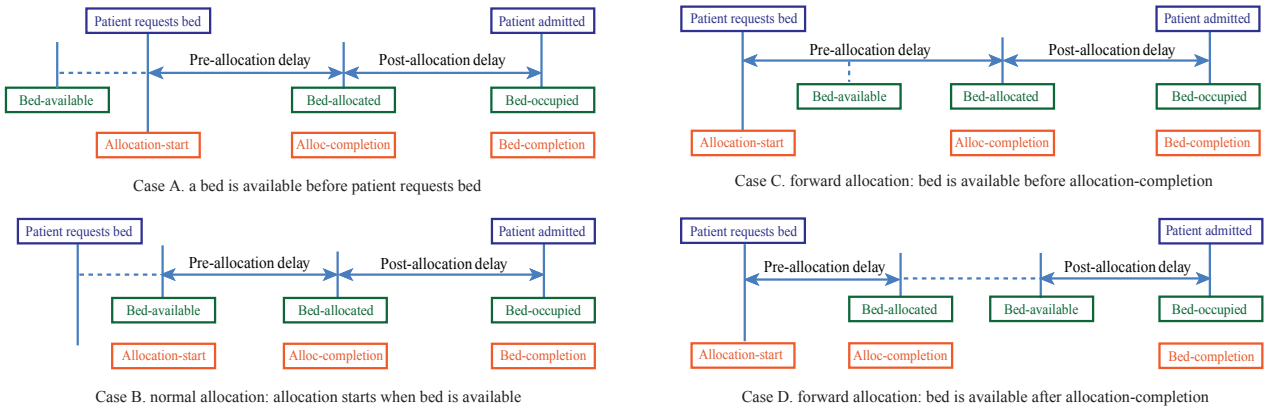


Figure 7 **Pre- and post-allocation delays under different scenarios.**

the empirical statistics in Sections 1 and 2. Note that in Section 4 of the online supplement [40] and on the MOH website [41], the waiting time of an ED-GW patient is defined as the duration between her bed-request time and departure time from ED, excluding the time between leaving ED and being admitted to a ward. The excluded duration is part of the post-allocation delay.

Allocating a newly arrived patient

We first assume a normal allocation. When a patient makes a bed-request, if a primary bed is available (Case A in Figure 7), the bed is selected for the patient and the bed allocation process starts immediately. When more than one primary pool has such a bed, a priority policy is used to decide which primary pool to choose from.

If no primary bed is available at the bed-request time, the patient waits in a buffer and is assigned with an overflow trigger time. The trigger time may depend on time of the day, admission source, and the specialty of the patient. An overflow policy dictates the specification of overflow trigger times (see Section 4.2). Before the overflow trigger time is reached, the patient waits for a primary bed and the bed allocation process starts at the availability of the primary bed. If no primary bed becomes available by the overflow trigger time and there is an overflow bed available, an overflow bed is selected following a certain policy and the allocation process starts at the overflow trigger time. If no overflow bed is available at the overflow trigger time, the patient continues to wait in the buffer for a bed from either a primary ward or an overflow ward. The allocation process starts when such a bed becomes available. In a normal allocation, the post-allocation delay starts at the allocation-completion time.

Now we assume a forward allocation. When a patient makes a bed-request and there is no primary bed available at the bed-request time, a not-yet-allocated bed is forward-allocated to her. To choose the bed, we follow a similar policy as above, i.e., if a primary bed is to be available before the patient's overflow trigger time is reached, the primary bed is chosen; otherwise, a primary or a overflow bed, whichever will be discharged first, is selected. The allocation process starts immediately at the bed-request time. When the patient finishes experiencing the pre-allocation

delay, the allocated bed may still be unavailable (Case D in Figure 7), in which case the patient waits until the forward-allocated bed becomes available, and the post-allocation delay starts at the bed-available time.

Allocating a newly-discharged bed

When a patient discharges from a bed and the bed has already been forward-allocated to a waiting patient, the post-allocation delay starts at this time if the pre-allocation delay has expired (Case D of Figure 7); otherwise, the post-allocation delay starts at the allocation-completion time (Case C of Figure 7). When the post-allocation expires, the patient is admitted into the bed, completing the bed-request.

When a patient is discharged from a bed and the bed has not been forward-allocated, if there is at least one patient who is eligible for this bed, one eligible patient is selected and the allocation process starts immediately. The eligible patients consist of both the primary patients whose waiting times are less than their overflow trigger times and the overflow patients whose waiting times are greater than their overflow trigger times. When there are multiple eligible waiting patients, a pre-specified priority rule is used to select one patient. When no eligible patient is waiting, the bed becomes available.

3.3. Service policies

A service policy governs all of the decisions regarding bed assignments at various decision time points. It has four components. These components are (i) how to pick a bed from a primary pool upon an arrival, (ii) how to pick a bed from a non-primary pool when a patient's overflow trigger time is reached; (iii) how to set an overflow trigger time; and (iv) how to pick a patient among a group of eligible patients upon a new discharge. We elaborate each component below.

Component (i) is a table that specifies the priority of primary pools. In general, dedicated pools have higher priorities than shared pools. Therefore, when seeking a primary bed for a patient, we start from the dedicated pools. If there is no dedicated bed free, we then search in shared pools.

Component (ii) is also a table that specifies the priority of non-primary pools to overflow a patient. The priority depends on the specialty of the patient to be overflowed. In general, pools that serve similar specialties have high priority. Shared pools have higher priority than dedicated pools. Both this table and the table in Component (i) are listed in Section 11.1 of the online supplement [40].

Component (iii) sets the overflow trigger time for those patients who have to wait due to the unavailability of any primary beds upon their bed-request times. When the overflow trigger time of such a patient is reached and she is still waiting for a bed, component (ii) is used to search for a non-primary bed. When no non-primary bed is available, the patient becomes an eligible patient. Section 4.2 specifies the details of assigning overflow trigger time in our model.

Component (iv) is a patient priority list, which will be given shortly below. It is used when a bed becomes free, and a patient from the pool of eligible patients needs to chosen to be admitted

into the bed. The priority list is built based on NUH's internal guideline [34] and our empirical observations. First, patients who have waited longer than their overflow trigger times have a higher priority than those who have not. This is aligned with NUH's goal of improving the 6-hour service level. Second, among the patients waiting longer than their overflow trigger times, those from the primary specialties have a higher priority than the ones from overflow specialties. Third, among patients from the same specialty, the ED-GW patients have a higher priority than ICU-GW and SDA patients, while ICU-GW and SDA have the same priority. This is based on the empirical observation that at NUH, ICU-GW and SDA patients have a much longer average waiting time than ED-GW patients. See Section 2.1 for our speculations. Also see [37] for a similar priority setting. Moreover, our model assumes that EL patients have the highest priority among all admission sources. We explain the reason of doing so in Section 3.5.1. Fourth, when patients are waiting in multiple buffers with the same priority or in a single buffer, we choose the patient with the longest waiting time.

3.4. Rationale to model transfer patients

This section explains our modeling of the transfer patients presented in Section 3.1. Recall that the term, general patients, refers to patients admitted to the general wards. At NUH, a general patient can be transferred from one ward to another, possibly multiple times, after initial admission. From NUH data, around 16% of all general patients have been transferred at least once after their initial admissions to general wards. Our model captures about half of these transfer patients, around 7% of total general patients, who are either ED-GW or EL source patients and are later transferred once or twice between general wards and ICU-type wards. Among the other half of transfer patients, most of them (around 7% of the total general patients) transfer from one general ward to another general ward, representing transfers from an overflow ward to a primary ward or from one primary ward to another primary ward that meets the financial expectation of the patient (e.g., upgrade to a high-end ward). Since the majority of these transfers occur in late afternoon (after the peak discharge period) and have little effect on the overall bed occupancy rate among general wards, we do not model them in our stochastic model. See Section 10.2 of the online supplement [40] for more details on the rational of not modeling them. The remaining transfer patients, about 2% of the total general patients, are transfers such as ICU-GW patients being returned to ICU-type wards, or ED-GW/EL patients who are transferred at least three times. These patients are ignored in our model since there are very few of them. We believe ignoring the half of transfer patients which we do not model has a limited effect on the main focus of this study as well as on our conclusions.

Among the transfer patients we choose to model, about 29% transfers once (from GW to ICU-type ward), and the remaining 71% makes two transfers (from GW to ICU-type ward to GW). Each transfer patient we model remains in a general ward prior to being transferred to an ICU-type ward. Empirical study shows that the LOS of this first stay, i.e. from initial admission to first transfer, is significantly shorter than the LOS of non-transfer patients. Thus, in our model

we divide ED-GW or EL patients into normal and transfer patients (see Table 1) to reflect the differentiation in LOS distributions. The discharge time and LOS of the ED-GW or EL transfer patient in our model correspond to the first transfer-out time and the first-stay LOS of the *real* transfer patient that we choose to model, respectively.

For a patient who is in the 71% two-transfer group at NUH, her second transfer is from an ICU-type ward to a general ward. For this patient, she stays in general wards twice, and the first and second ward may be different. The preceding paragraph details how to model her first stay. To model the second stay of this real patient in a general ward, we create a pseudo-patient. The admission time of this pseudo-patient corresponds to the transfer-in time of the real patient (from an ICU-type ward to a general ward), and the discharge time of this pseudo-patient corresponds to the final discharge time (from the general ward) of the real patient. Thus, the LOS of the pseudo-patient corresponds to the number of nights in the second stay of the real patient. Although the real patients are either from ED-GW source or from EL source, we treat the pseudo-patients as ICU-GW patients because the admission process and admission time distribution of the pseudo-patients are close to those of the real ICU-GW patients. However, empirical analysis shows that the LOS distribution of the pseudo-patients is different from that of the real ICU-GW patients in the same specialty. To differentiate these two groups of ICU-GW patients, we classify the ICU-GW patients into newly-admitted and re-admitted patients, where the latter refers to these pseudo-patients. See Section 4.3 for details on estimating the LOS and discharge distributions for all transfer patients in the model.

Our model is a parallel-server-pool system (server pools corresponding to general wards) with a single-pass routing structure. In particular, we do not model ICU-type wards and patient flows within ICU-type wards in our system. Without creating pseudo-patients, the second transfer flow from ICU-type wards to general wards would be lost in our model.

3.5. Summary of inputs to the model

We summarize the inputs needed to populate the model. Details of certain critical modeling elements are further elaborated in Section 4.

3.5.1. Patient arrivals, type and class designations

As shown in Figure 6, patient arrivals to our model derive from four sources. For each source, the arrival rate is time-dependent. We assume the arrival rate function is periodic with one day as the period. For ED-GW, ICU-GW, and SDA patients, we estimate their hourly bed-request rates empirically (use the bed-request time stamps) and use these estimations as the arrival rates in our model. For EL patients, their arrivals are pre-scheduled. NUH has their admission times but lacks meaningful records on bed-request times for these patients. Thus, we empirically estimate their hourly *admission* rate and use the estimation as their hourly arrival rates in our model. We assign EL patients the highest priority and set their allocation delays to be zero. In this way, the waiting times of EL patients in our model are negligible, and hence their admission times are close to their

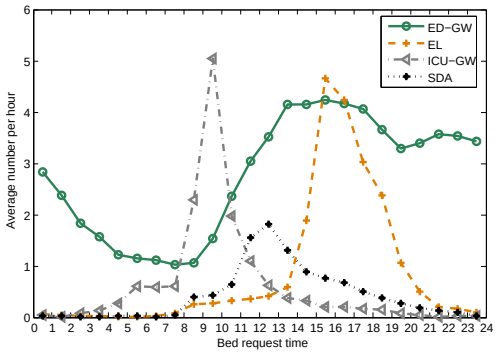


Figure 8 **Hourly arrival rate for each admission source (estimated from Period 1 data). The daily arrival rate of each source is close to its daily admission rate shown in Figure 3a, except for ICU-GW source since re-admitted patients are included here.**

bed-request times. Therefore, our model input for the EL patient bed-request rate is reasonable. Figure 8 shows the estimated hourly arrival rates for the four sources, which we use as model inputs.

In our model, we assume that the four arrival processes are periodic time-nonhomogeneous Poisson processes (with one day as the period). Section 6.2 of the online supplement [40] presents a detailed study on testing the assumption of time-nonhomogeneous Poisson for the bed-request process of ED-GW patents. Following a statistical procedure proposed in [5], we perform 30 tests, one for each month in the two periods, on the empirical bed-request times. Among the 30 tests, 24 of them do not reject the hypothesis that the bed-request process of ED-GW patients follows a time-nonhomogeneous Poisson process with piecewise-constant arrival rates (see Table 8 there). Therefore, it is reasonable to assume that the bed-request process for ED-GW patients is nonhomogeneous Poisson. However, Figure 15 of [40] suggests the bed-request process is *not a periodic* Poisson process with either one day or one week as a period. In particular, the empirical CV of the daily arrival rate for each day of week is much higher than 1, the theoretical CV under the Poisson assumption. It is conjectured that the high variability comes from the seasonality of bed-requests and the overall increasing trend in the bed demand (see more discussions in Section 6.2 of [40]). For bed-request processes from SDA and ICU-GW sources and the EL admission processes, we have done similar tests. These tests rejects the hypothesis that they are time-nonhomogeneous Poisson.

Despite the lack of strong empirical support, following a standard practice in literature, we assume that the four arrival processes are time-nonhomogeneous Poisson for convenience. We further assume that each non-homogenous Poisson process is periodic with one day as a period. The latter assumption is appropriate given our focus is to stabilize the waiting time statistics within a day and to evaluate operational policies (e.g., discharge policy) under a stable environment that includes a daily stationary arrival process. Setting one week as a period is another reasonable choice, and we leave this extension to a future study to capture the day-of-week phenomenon. We also point out that building new arrival process models is an active, challenging research problem; see recent work of Glynn [14].

p	Surg	Card	Med	Ortho	Onco
ED-GW	4.58%	11.52%	4.78%	9.42%	5.69%
EL	23.46%	39.95%	4.53%	17.04%	6.01%
ICU-GW	45.10%	43.86%	16.98%	79.69%	39.86%

Table 2 Estimated value for the parameter p of the Bernoulli distribution to determine patient classes. For ED-GW and EL patient types, p represents the probability of being a transfer patient; for ICU-GW, p represents the probability of being a re-admitted patient. Parameters for specialties belonging to the Medicine cluster (Gen Med, Gastro-Endo, Neuro, Renal, Respi) are estimated together due to the limited number of data points, and we use Med to represent this group.

When an arrival from an admission source occurs, we randomly assign the arriving patient to one of the nine specialties following an hour- and admission-source-dependent empirical distribution. There are a total of $24 \times 4 = 96$ distributions, and we obtain each of them from the proportion of patients from each specialty out of the total arrivals of that source in that hour. Figure 3b plots the daily distributions of specialties and admission sources.

As discussed in Section 3.1, the chosen speciality necessarily determines the patient type. Within each type, we choose a class for the patient at the time of admission, following a distribution that is type and admission-period (if applicable) dependent. These distributions are specified in Section 3.5.3. Since a thinning of a Poisson process is still Poisson [39], the arrival process of each of the 36 patient types and of each class is Poisson in our model.

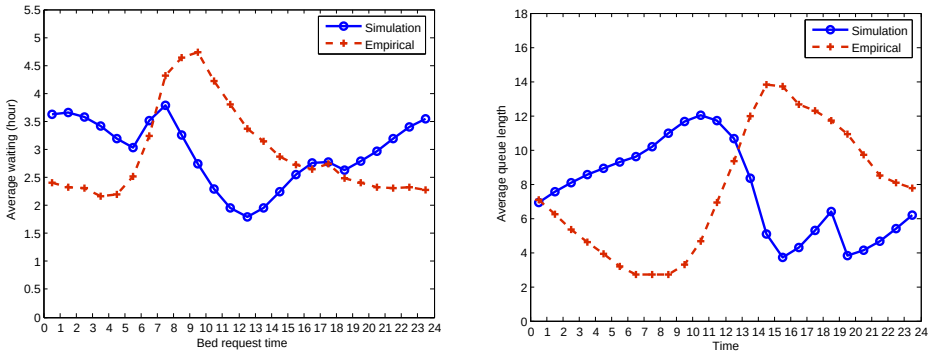
3.5.2. Server pools and service policy

Table 17 in Section 11.1 of the online supplement [40] lists the number of servers and the primary specialties for each server pool. Table 18 of the online supplement specifies the priority table for components (i) and (ii) of the service policy discussed in Section 3.3. Component (iii) of the service policy is elaborated in Section 4.2.

3.5.3. Allocation delays and service time

We assume that pre- and post-allocation delays follow log-normal distributions with time-dependent means. We assume all patients have the same pre- and post-allocation delay distributions. See Section 4.1 below for more details on the allocation delay distributions and the time-dependent means.

As mentioned in the introduction, we do not model service time as an exogenous variable. Instead, service time depends on the two exogenous random variables, LOS and discharge time. Both of these variables are class-dependent; see more details in Section 4.3. The patient's class under a given type is determined upon her admission time. For an ED-GW patient, the admission period (AM or PM) is known at that time. For an ED-GW patient in either admission period or an EL patient, we randomly assign her as normal (non-transfer) or transfer following a Bernoulli distribution. For an ICU-GW patient, upon the admission time, we assign her as newly-admitted or re-admitted following a Bernoulli distribution. The parameters of these Bernoulli distributions depend on specialty and their empirical estimations are given in Table 2.



(a) Average waiting time (b) Average queue length

Figure 9 **Simulation estimates of the hourly average waiting time and hourly average queue length for ED-GW patients: not modeling the two allocation delays. Empirical estimates are from Period 1 data.**

4. Model elements that are important for model fidelity

In this section, we identify a few key elements that are important for the fidelity of our model. These elements include pre- and post-allocation delays that create additional delay during patient's admission and discharge, non-iid service time model, and a time-dependent dynamic overflow policy. Missing or improperly modeling any of these elements will make the model irrelevant. We also specify additional details of the model inputs that are included in Section 3.5.

4.1. Pre- and post-allocation delays

As noted in the introduction, a key feature of our model is to explicitly model operational delays in the bed allocation and admission processes of a patient. To show the necessity of modeling allocation delays, Figure 9 compares the simulation and empirical estimates of the hourly average waiting time and hourly average queue length for ED-GW patients. In the simulation setting, *no* allocation delays are modeled. We can see that the hourly performance curves from simulation are completely different from the empirical estimates. In particular, note that the blue curve in Figure 9b, which shows a rapid drop in the simulated average queue length between 11am and 3pm, contrasts sharply with the empirical (red) curve, which drops slowly after 2pm. The main reason for the rapid drop in the blue curve is that in Period 1, between 11am and 3pm, the discharge rate increases in each hour until reaching the peak at 2-3pm (see Figure 2), and a waiting patient in the simulation is admitted into service immediately once a discharge occurs. Thus, Figure 9 suggests the existence of extra delays after bed discharges. In this simulation study, to make the daily average waiting time still comparable to the empirical estimate (2.82 hours), we decrease the numbers of servers listed in Table 17 of the online supplement, while keeping all other settings the same as those used to produce Figure 5.

In this section, we focus on estimating allocation delays for ED-GW patients. We first explain how to model allocation delays for other patients. We assume the allocation delays of the EL patients to be zero in the model, and we explained the rational of doing so in Section 3.5.1. For ICU-GW and SDA patients, we do not have good time stamps to estimate of their pre- and post-allocation delays

reliably. We simply assume their allocation delays are similar to the allocation delays experienced by ED-GW patients. Therefore, allocation delays of ICU-GW and SDA patients are drawn from the same distributions that are used to generate allocation delays for ED-GW patients. Sensitivity analysis shows that a moderate amount of change to the allocation delay distributions of ICU-GW and SDA patients will not affect the overall performance of ED-GW patients.

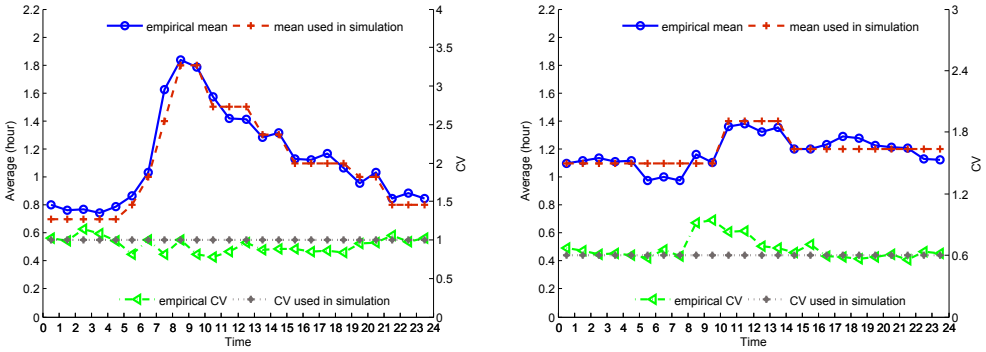
Estimate pre-allocation delay

In NUH data set, at the bed-request time of an ED-GW patient, either (i) the allocated bed is already available for the patient, or (ii) the bed is not available and still occupied by another patient. We select a subset of case (i) patients in the data set to estimate the pre-allocation delay distribution. The subset consists of case (i) patients whose allocated beds are from their primary wards. By selecting this group of patients, we try to minimize the influence of bed shortage and specialty mismatch on pre-allocation delay so that our estimation can reflect the minimum time needed for BMU agents to allocate a bed. For the selected patients, their pre-allocation delays start from the bed-request times, and end at the allocation-completion times.

To motivate a probabilistic model for pre-allocation delays, Section 9.2 of the online supplement [40] plots the histograms of the pre-allocation delays and the fitting results for the log-transformed data points against normal distributions. Sub-groups are created to account for the time-dependent feature of pre-allocation delays, which we will emphasize in the following paragraphs. These empirical observations suggest that using log-normal distribution is a good starting point for modeling the pre-allocation delay. Thus, our model assumes the pre-allocation initiated within each hour of a day to be a random variable that follows a log-normal distribution. The mean and variance of the log-normal distribution depends on the initiation hour (i.e., the hour when the pre-allocation starts).

To completely specify the pre-allocation distributions, we need to specify both the mean and coefficient of variation (CV) of the pre-allocation delay as a function of the delay initiation hour. Figure 10a shows the plots of the mean (blue curve) and CV (green curve) from empirical estimates. In our simulation, we use the red and grey curves as the inputs for the time-dependent mean and CV, respectively. These two curves are slightly smoother than the corresponding empirical curves, which have random noises since the sample sizes in certain time intervals are small, particularly between 8am and 10am. The red and grey curves are well within the 95% confidence interval of the empirical curves. We report that as long as the means and CVs for pre-allocation delays are taken from the red and grey curves in Figure 10a, the hourly waiting time performances are not sensitive to the choice of pre-allocation delay distributions. This observation also applies to post-allocation delays that will be discussed in the next section.

Figure 10a clearly demonstrates a time-dependent feature of the pre-allocation delay. The average delays are longer if the delay initiation time is in the morning. We speculate that the reason is because ward nurses/physicians are busy with morning rounds, and have less time to accept new patients. Therefore, it takes BMU longer time to search and negotiate for beds in the morning.



(a) Pre-allocation delay (b) Post-allocation delay

Figure 10 **Average and CV of estimated pre- and post-allocation delays with respect to the delay initiation hour. Left vertical axis is for the average; right vertical axis is for the CV. The scale of the right vertical axis is deliberately chosen to be large, so that the four curves are not crossed over.**

Estimate post-allocation delay

We select all ED-GW patients in NUH data set to estimate the post-allocation delay distribution. For a selected patient, her post-allocation delay starts at the allocation-completion time if the allocated bed is available at the time of allocation, otherwise it starts at the bed-available time (see Figure 7). The post-allocation delay ends when the patient occupies the bed at the bed-request-completion time.

Similar to the empirical analysis for pre-allocation delays, Section 9.2 of the online supplement [40] plots the histograms for post-allocation delays. The histograms and distributional fitting results in [40] again suggest that we use a log-normal random variable to model the post-allocation delays that are initiated in each hour. The blue and green curves in Figure 10b show the empirical estimation for the average and CV of post-allocation delay with respect to the delay initiation time, respectively. In the simulation, we use the red and grey curves in Figure 10b as the input for the mean and CV in each hour, respectively. We slightly adjust the simulation input based on the empirical estimate for the same reason we adjusted it in the pre-allocation delay.

In Figure 10b, we again observe a time-dependent feature. The longer averages in the morning may mainly stem from the ED side. The ED at NUH is usually congested in late mornings, so it is likely that ED physicians and nurses are busy with newly arrived patients and have less time to discharge and transfer admitted patients to general wards.

Estimate the normal allocation probabilities $p(t)$

Recall from Section 3.2 that in the model, when a patient makes a bed-request at time t and there is no primary bed available at the time, we assume with probability $p(t)$ the allocation for the patient is a normal allocation, meaning this patient will wait until a bed is available before the

allocation process starts. Unfortunately, the NUH data sets do not have accurate time stamps to let us estimate $p(t)$ reliably. In all the simulation runs in this paper, we choose

$$p(t) = \begin{cases} 0 & t \in (0,6], \\ .25 & t \in (6,8], \\ 1 & t \in (8,12], \\ .75 & t \in (12,14], \\ .5 & t \in (14,20], \\ 0 & t \in (20,24]. \end{cases} \tag{1}$$

In the next four paragraphs, we explain the rationales for using $p(t)$ given in (1).

First, the choice of $p(t) = 1$ between 8am and 12 noon is consistent with the current practice at NUH. In order to do a forward allocation, the planned discharge information should be available. Most wards are doing the morning rounds at about 9-11am, and nurses would only know which patients will be discharged after finishing the rounds. Thus, BMU typically receives the planned discharge information when it is close to noon.

Second, between 2pm and 8pm, for each hour i, we use $\hat{p}(i)$ to estimate $p(t)$ for $t \in (i, i+1]$, where

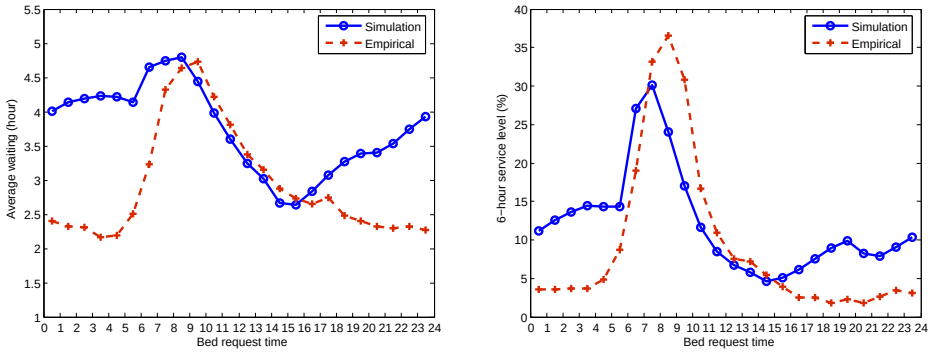
$$\hat{p}(i) = \frac{\#\text{ of patients whose allocation-completion time } > \text{ bed-available time}}{\text{total}}. \tag{2}$$

Here the total patient group consists of all ED-GW patients (in the NUH data) whose bed-request time falls within that hour and whose allocated bed is not available at the bed-request time. The patients included in the numerator correspond approximately to normal allocations in the model (in this time interval). Figure 35 in the online supplement [40] shows that, between 2pm and 8pm, the ratio $\hat{p}(i)$ in (2) fluctuates near the $(40\%, 50\%)$ range. Based on this observation, we set $p(t) = .5$ between 2pm and 8pm. Section 9.2 of the online supplement [40] explains why patients in the numerator correspond to normal allocations and why we can use (2) to estimate $p(t)$ in certain time intervals.

Third, NUH data also shows that, between 8pm and 6am the next day, there are very few (fewer than 15 each hour) normal allocations, suggesting $p(t)$ is close to zero. Therefore, we set $p(t) = 0$ in this time interval.

Fourth, during each of the remaining time intervals of a day, $(6,8]$ or $(12,2]$, we estimate $p(t)$ by interpolating its values in its neighboring intervals to avoid the sudden changes of $p(t)$. The actual values of $p(t)$ in these two intervals are obtained by trial-and-error so that red and blue curves in Figure 5 are as close as possible.

We realize that, despite of our best effort, our choice of $p(t)$ using (1) is still ad hoc. In Section 11.3 of the online supplement [40], we conduct a sensitivity analysis of the choice of $p(t)$. The analysis shows that the conclusions in Section 5 are robust. In our current model, the allocation-start time is either at bed-request time or at bed-available time. An alternative model is to randomly assign it following a certain distribution. We leave this extension to a future study.



(a) Average waiting time (b) 6-hour service level

Figure 11 **Hourly average waiting time statistics from simulation output: static overflow policies with fixed overflow trigger time $T = 4.0$ hours.**

4.2. A dynamic overflow policy

As mentioned in Section 3.2, when a patient's waiting time reaches the pre-assigned overflow trigger time T, overflow to a non-primary pool may occur. In our model, we call the mechanism to determine this overflow trigger time T an *overflow policy*.

At NUH, there is a general guideline [34] on when and how to overflow a patient. Consistent with this guideline, empirical evidence [42] suggests that the hospital overflows patients more aggressively during late night and early morning (before 7am). That is, NUH would overflow a patient almost immediately upon finding that no primary bed is available. The reason is that few discharges happen in this time period, so there is little chance that a primary bed will become available in the next few hours. Thus, there is no need to let the patient wait for another hour. In contrast, during other times, the hospital tends to be more conservative, and allows a patient to wait some time prior to overflow in anticipation that a primary bed may become available soon. In this way, NUH has better control on the overflow proportion, another important performance metric being monitored (see Section 2.3). The preceding discussion suggests that the trigger time T should depend on the bed-request time. It is reasonable to assume that T is low when a bed-request occurs during late night or early morning, and high during other times.

Based on these observations, we use a simple dynamic overflow policy in our model: when a patient requests a bed from 7am to 7pm, the overflow trigger time T is set to be $t_2 = 5.0$ hours, and for bed-requests in all other time periods, T is set to be $t_1 = 0.2$ hour. From Figure 5 we can see that using the simple dynamic overflow policy with this set of parameter values can reproduce the hourly waiting time performances. In contrast, Figure 11 compares the same empirical estimate with the simulation estimate from our model but with a static overflow trigger time $T = 4.0$ hours. Clearly, the model with static trigger time fails to capture the dynamics in NUH inpatient operations.

We choose 7am and 7pm as the starting and ending point to adopt the long overflow trigger time, respectively. This choice is based on observations from [42] and the practice at NUH. 7pm to 7am the next day is the night-shift period at NUH. A nurse manager is in charge to deal with all

bed-requests in this period. She has the authority to overflow patients without negotiation. The values of t_1 and t_2 are obtained through trial-and-error so that the simulation output curves in Figure 5 are as close to the empirical curves in the figure as possible. It is important to note that overflow decisions are very complicated [42], sometimes subjective, in practice. There is no data available for us to get an accurate estimation of the overflow trigger time. Thus, our proposed dynamic policy is an approximation of the real situation. Other variants of the overflow policy are possible, e.g., triggering an overflow event when the number of waiting patients exceeds a specified threshold, selecting T based on the remaining service times of patients who are in service. We leave this extension for future study.

4.3. Length of stay and non-iid service times

The service time of a patient is the duration between the admission time and the discharge time. We use day as the time unit unless specified otherwise. Clearly, the service times of patients are random. A number of factors affect the service time. Not surprisingly, the medical diagnosis of the patient affects her service time. Moreover, resource constraints such as staffing levels and operational policies can also affect the admission and discharge timing, and thus affect service times. To separate these two different sources of influence on service times, we adopt the following decomposition for the service time S of a patient in our model:

$$S = \mathrm{LOS} + h_{\mathrm{dis}} - h_{\mathrm{adm}}. \tag{3}$$

Here, LOS stands for length of stay and is equal to the number of midnights that the patient spends in a general ward, or equivalently, day of discharge minus day of admission, and h_{dis} and h_{adm} stand for the hour of the day when the patient is admitted and discharged, respectively. The hour of the day is between 0 and 1, with midnight being 0 day and 12pm (noon) being .5 day. For a patient who is discharged on the same day of admission, our definition of her LOS is equal to 0, whereas in medical literature it is set to 1 [7, 20]. The latter adjustment is necessary, for example, for accounting and cost recovery purposes.

After an extensive empirical study (see Sections 7 and 8 of the online supplement [40]), we make the following assumptions for the service times.

1. We assume that h_{dis} is independent of LOS and of h_{adm}; this assumption is reasonable because we believe LOS captures the amount of time that a patient *needs* to spend in a ward due to medical reasons, whereas discharge hour h_{dis} clearly depends on the discharge patterns, which are mainly the results of scheduling and behaviors of medical staff. Section 8.5 of [40] also provides some empirical evidence for this assumption.

2. LOS distributions are class dependent. Definition of class can be found in Section 3.1. Table 3, which lists the average LOS for each class of non-transfer patients, clearly shows the dependency on admission source and specialty. Transfer patients have a different set of LOS distributions from their non-transfer counterparts, and we discuss estimating their LOS distributions at the end of this section.

Cluster	ED-GW(AM)	ED-GW(PM)	EL	ICU-GW	SDA
Surg	2.36 (2.93)	3.27 (3.43)	4.55 (6.55)	9.58 (12.60)	2.59 (4.72)
Card	2.95 (3.75)	3.83 (3.93)	4.15 (5.08)	5.22 (6.78)	2.55 (3.38)
Gen Med	3.94 (4.76)	5.25 (5.87)	5.32 (5.79)	10.43 (18.43)	3.17 (2.62)
Ortho	5.45 (8.22)	6.04 (7.04)	6.27 (6.19)	10.82 (13.32)	3.41 (4.32)
Gastro	3.32 (3.91)	4.48 (4.47)	3.70 (4.39)	8.33 (12.25)	3.24 (3.99)
Onco	5.93 (7.58)	7.03 (7.14)	6.45 (7.95)	8.62 (9.02)	4.10 (4.18)
Neuro	3.23 (5.22)	4.07 (4.69)	4.06 (4.69)	7.56 (7.67)	2.59 (2.40)
Renal	5.75 (6.55)	6.51 (6.90)	5.70 (6.20)	10.22 (12.91)	2.08 (1.16)
Respi	3.21 (5.10)	4.29 (4.26)	4.45 (6.27)	7.86 (10.71)	2.33 (3.33)
All	3.70 (5.25)	4.78 (5.45)	5.17 (6.47)	7.59 (10.82)	2.84 (4.29)

Table 3 Average LOS (in days) for patients in each specialty from different admission sources; only non-transfer patients are displayed here. Period 1 data is used, and number in the bracket is the corresponding standard deviation (std).

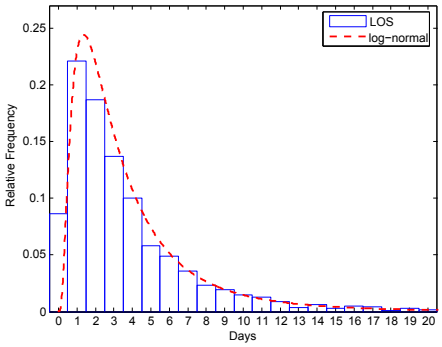


Figure 12 **LOS of ED-GW patients from General Medicine. Only non-transfer AM patients in Period 1 are included. The LOS distribution can be fitted with a log-normal distribution (mean 3.94, std 4).**

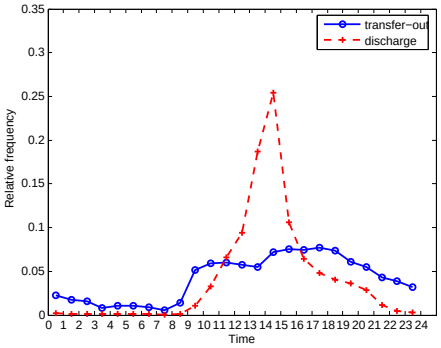
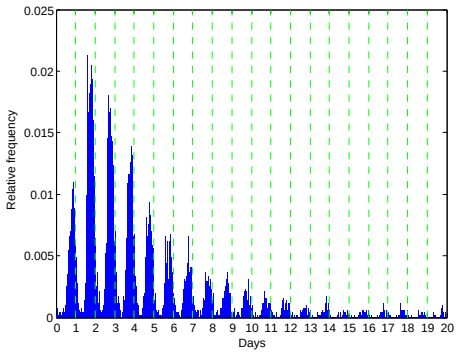
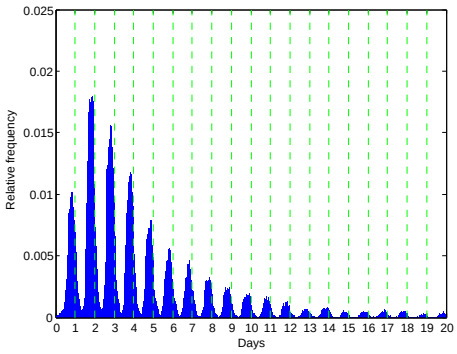


Figure 13 **Discharge distributions for transfer patients from ED-GW and EL sources and for non-transfer patients. The blue curve is estimated from the combined data, and the red curve is estimated from Period 1 data.**



(a) Empirical service time distribution for Period 1; each green dashed line corresponds to a 24-hour increment



(b) Service time distribution from simulation output; LOS and discharge distributions in the simulation are empirically estimated from Period 1 data

Figure 14 **Service time distributions, in hourly resolution, for General Medicine patients (normal-PM class).**

3. For each class of patients, we assume the LOS to be iid following an empirical distribution. See Section 7.3 of the online supplement [40] for plots of the empirical LOS distributions for each

class of non-transfer patients. Each of these empirical distributions can be fitted with a log-normal distribution, although this fact is not used in this paper. See Figure 12 for an example of fitting the empirical distributions.

4. We assume the discharge hours h_{dis} among all non-transfer patients and re-admitted ICU-GW patients, across all types, are iid random variables following the same discharge distribution. The discharge distributions for Period 1 and Period 2 are different, and they are plotted in Figure 2.

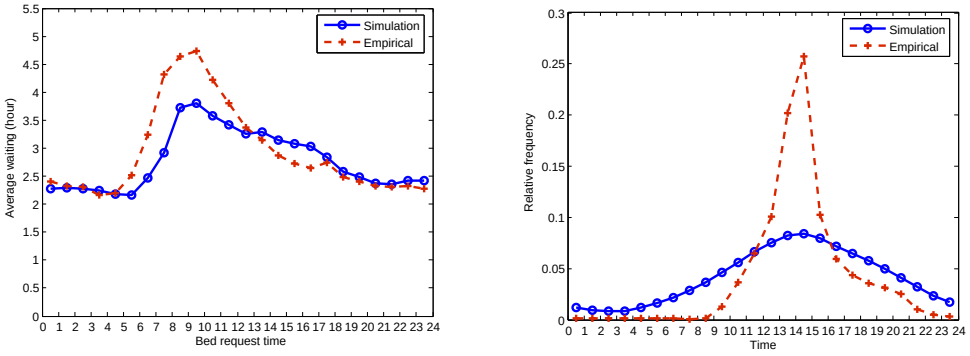
Discharge hour distribution among all transfer patients from ED-GW and EL sources, across all types, follows from a different discharge distribution. This discharge distribution is empirically estimated from the first transfer-out times (from a general ward to an ICU-type ward) of all the real transfer patients we model, and is plotted in Figure 13. We do not observe significant difference between the two periods.

5. We assume LOS are independent among all general patients, i.e., no dependency between any classes.

We want to emphasize that LOS distributions are admission-period dependent for non-transfer ED-GW patients. Table 3 shows that for each specialty of ED-GW patients (non-transfer), a before-noon admission (AM) patient on average spends one day less than an after-noon admission (PM) patient. We speculate the reason might be that the rest of the admission day can be used for further medical diagnosis for AM patients, but not for PM patients. For patients from the other three admission sources, we do not assume their LOS to be admission-period dependent because there are very few AM patients. Indeed, EL, ICU-GW and SDA have only 4%, 6%, and 8% of before-noon admissions, respectively. Section 7.2 of the online supplement [40] contains more empirical findings on the AM/PM differentiation.

Empirical analysis in [40] shows that there is no significant difference between two periods for the LOS distributions. The simulation output in Figure 5 is generated from our proposed service time model using the empirical LOS and discharge distributions in Period 1. Figure 14 illustrates that our proposed service time model can produce the service time distributions that resemble empirical distributions. We observe a *clustering* phenomenon from both Figures 14a and 14b. Section 8.2 of the online supplement [40] discusses this clustering phenomenon. Armony et al. [1] discover the same phenomenon in an Israeli hospital when plotting service time distributions using the same hourly resolution.

Under our service time model assumptions (1)-(5), for a class of patients, their LOS form a sequence of iid random variables and their discharge hours h_{dis} form another sequence of iid random variables, and these two sequences are independent. But admission hours h_{adm} of these patients are ordered, so they cannot be iid. It follows from Equation (3) that the service times are *endogenous* and therefore *not* iid. We believe that to replicate the hourly performance curves, the endogenous service time model is an important component for modeling inpatient flow operations. This model is contrary to the exogenous service time model often used in the queueing literature. We compare



(a) Hourly average waiting time curve (b) Discharge distribution

Figure 15 **Simulation output from using an iid service time model.**

an iid exogenous service time model with our proposed non-iid model. The iid model assumes the service time S to be the sum of two independent random variables: an integer variable corresponding to the floor of service time $\lfloor S \rfloor$, and a residual variable corresponding to $(S - \lfloor S \rfloor)$. For patients from the same class, we assume their integer parts and residual parts each form an iid sequence with an empirical distribution. Since the two sequences are independent, the service times are iid. See Section 8 of the online supplement [40] for empirical evidence supporting the independence assumption between the integer and residual variables, as well as plots of their distributions. Even though this iid exogenous service time model can reproduce the service time distributions such as the one in Figure 14a, it is not able to reproduce the discharge distribution and the hourly waiting time statistics; see the simulation output in Figure 15 for an illustration. Therefore, we advocate the endogenous service time model for inpatient flow management.

Estimating LOS distributions for transfer patients

Section 3.4 explains the rationale to model patient transfer. The transfer patients we model are ED-GW or EL source patients at NUH who transfer once or twice between general wards and ICU-type wards after the initial admission. For each of the modeled *real* patients, her first visit to a general ward starts from the initial admission time and ends at the first transfer-out time to a ICU-type ward. If she transfers twice, her second visit to a general ward starts from the transfer-in time (from ICU to GW) and ends at the final discharge time.

We use the first- and second-visit LOS of these real patients, i.e., number of nights in the corresponding visit, to empirically estimate the LOS distribution for transfer patients in our model. Specifically, we use the first-visit LOS of the real ED-GW patients with admission time before and after noon to estimate the LOS distribution for transfer-AM and transfer-PM class patients (under ED-GW source), respectively. We use the first-visit LOS of the real EL patients to estimate the LOS distribution for EL-transfer class patients. The second-visit LOS of all modeled patients who transfer twice is used to estimate the LOS distribution for re-admitted ICU-GW patients. See Section 10.3 of the online supplement [40] for more details on estimating the LOS distributions for transfer patients and the corresponding plots.

5. Factors that impact hospital performance

There is no existing analytical tool to analyze either exactly or approximately the proposed stochastic processing network model with all the key elements described in Section 4. In this paper, we analyze the stochastic model via computer simulations. The simulation code is written in `C++` language. The inputs to the simulation model are either estimated or motivated from the empirical analysis that we have conducted for the NUH inpatient department; they are summarized in Section 3.5. For each simulation run, we simulate the model for a total of 10^6 days, and divide the simulation output into 10 batches. The performance measures are calculated from averaging the last 9 batches, with the first batch discarded to eliminate transient effects. Unless otherwise specified, all simulation estimates in this section are from simulation runs under this setting. In Section 5.1, we establish a baseline scenario that corresponds to NUH Period 1 operation. Simulation outputs from this baseline scenario match the empirical performances in Period 1. In Section 5.2, we show that the early discharge in Period 2 has little impact in achieving time-stable waiting time statistics. In Section 5.3, we show that a hypothetical Period 3 policy can stabilize the waiting time statistics. This policy is not completely practical, but can serve as an inspirational goal for hospital mangers to aim at. The Period 3 policy requires improvement in both the discharge distribution and allocation delays. In Section 5.4, we show that simultaneous improvement is necessary to achieve stable waiting time statistics. We also demonstrate that increasing bed capacity within a certain range does not necessarily stabilize the waiting time statistics.

5.1. The baseline scenario

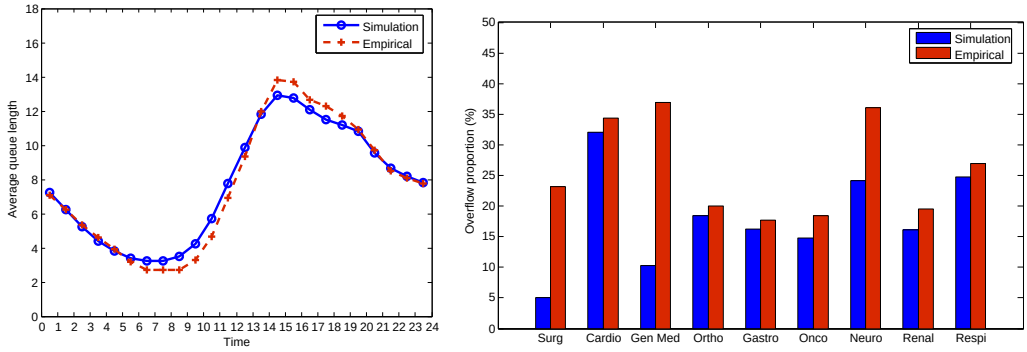
In the baseline scenario, the model inputs have the same setting as summarized in Section 3.5. Basically, for all distributions estimated from the empirical data, including LOS and discharge distributions, we use Period 1 data unless specified otherwise. From simulating the baseline scenario, the overall average waiting time for ED-GW patients is 2.81 hours and the 6-hour service level is 6.23%. Figure 5 in Section 3 shows that the simulation estimates match the empirical estimates of the time-varying waiting time statistics for all ED-GW patients. Table 4 compares the simulation estimates with the empirical estimates of the average waiting time and the 6-hour service level for each specialty. We can see that the waiting time statistics, even at specialty level, can be matched by our simulation.

Besides waiting time, we can also calibrate other key performance measures. The overall utilization rate is 90.2% from simulation, a little bit higher than the empirical utilization 88.0% in Period 1. Figure 16a plots the hourly average queue length for all ED-GW patients from simulation and empirical estimates.

Given that the simulation estimates of major performance measures under the baseline scenario are close to the empirical estimates from Period 1 data, we are confident that our baseline model is properly calibrated, at least for predicting the waiting time statistics. In the next few sections, using the proposed model we start from evaluating the impact of the early discharge policy implemented

	avg waiting time (hour)		6-h ser level (%)	
Specialty	simulation	empirical	simulation	empirical
Surg	2.61 (0.01)	2.61	4.58 (0.06)	5.45
Card	2.94 (0.01)	3.08	6.53 (0.08)	8.36
Gen Med	2.71 (0.01)	2.64	5.21 (0.06)	4.79
Ortho	2.70 (0.01)	2.79	5.01 (0.05)	5.84
Gastro	2.91 (0.01)	2.97	8.54 (0.09)	7.64
Onco	2.86 (0.01)	2.96	7.61 (0.09)	8.15
Neuro	2.82 (0.01)	2.81	6.34 (0.07)	5.93
Renal	3.33 (0.01)	3.41	11.49 (0.12)	11.60
Respi	2.80 (0.01)	2.77	6.08 (0.08)	5.50
All	**2.81** (0.01)	**2.82**	**6.23** (0.06)	**6.52**

Table 4 Simulation and empirical estimates of waiting time statistics for ED-GW patients from each specialty. The simulation estimates are from simulating the baseline scenario; the number in the parentheses is the standard deviation of the corresponding value. The empirical estimates are from Period 1 data.

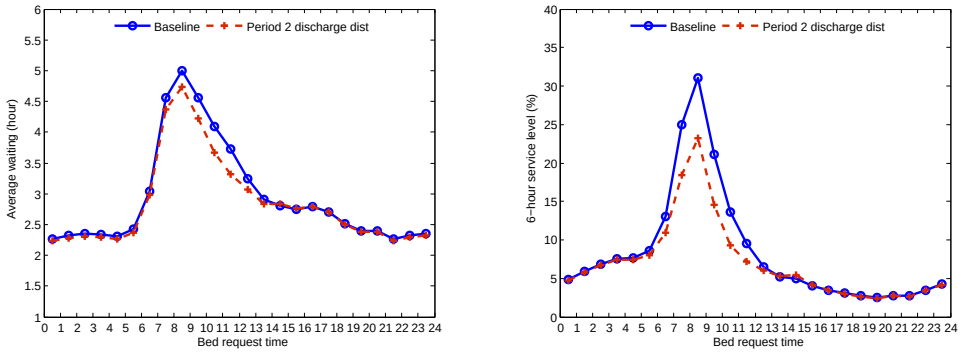


(a) Average queue length (b) Overflow proportion

Figure 16 **Simulation output compares with empirical estimates: hourly average queue length and overflow proportion (Period 1).**

at NUH in Period 2. We then do what-if analyses to identify factors that impact the hospital performances and propose policies under which waiting time statistics can be stabilized.

We point out that our model cannot perfectly replicate the overflow proportion. Although the simulated overflow proportions for most specialties are close to their empirical counterparts (see Figure 16b), the baseline simulation underestimates the overflow proportions for Surgery, General Medicine, and Neurology specialties. The underestimation in these three specialties leads to an overall underestimation of overflow proportion across all specialties (17.04% in the baseline versus 26.95% from Period 1 data). Section 11.4 of the online supplement [40] explains the challenges in calibrating overflow proportions. Finally, there are certain performance measures that we choose not to calibrate in the model, including the waiting time statistics for ICU-GW and SDA patients. As mentioned, the waiting time statistics for these patients are not the focus of this paper. Moreover, sensitivity analysis shows that whether we can accurately replicate their waiting times has little impact on the waiting time statistics of ED-GW patients. Readers are referred to Section 6 for more discussions.



(a) Average waiting time (b) 6-hour service level

Figure 17 **Comparing hourly waiting time statistics under the baseline scenario and scenario with Period 2 discharge distribution.**

5.2. NUH early discharge has little impact on stabilizing waiting time statistics

To evaluate the impact of early discharge policy on waiting time statistics, we simulate a scenario with the same inputs as in the baseline scenario, but using the discharge distribution estimated from Period 2 data (i.e., using the red curve in Figure 2 to replace the blue curve). Figure 17 compares the simulation estimates of hourly waiting time statistics with the baseline scenario. From Figure 17a, the hourly average waiting times show little difference in the two scenarios. From Figure 17b, the hourly 6-hour service level exhibits some reductions for bed-requests between 7am and 11am, e.g, the peak value is now 23% compared to 31% in the baseline scenario, but the values for other hours are almost identical in both scenarios. Not surprisingly, other performance measures from these two scenarios are almost identical. The overall average waiting time under this early discharge scenario is 2.73 hours, a 4-minute reduction, versus 2.81 hours in the baseline scenario. The 6-hour service level is 5.50% versus 6.23% in the baseline scenario. The overall overflow proportion is 16.91%, not significant different from the baseline value 17.04%.

We conclude that early discharge policy, implemented at the level that NUH has achieved by December 2009, has limited impact on improving the daily waiting time statistics for ED-GW patients and the overall overflow proportions. Moreover, Period 2 discharge policy alone cannot stabilize the waiting time statistics throughout the day. This discovery is consistent with the findings from a study at a large U.S. hospital [16].

Readers should be aware that, like many public service systems in Singapore, NUH has been constantly seeking operational improvement. Even in Period 1, NUH manages the discharge planning in a more efficient way than many hospitals around the world. For example, [1] reports that an Israeli hospital has its peak discharge time between 4pm and 5pm, two hours later than the peak discharge time in Period 1 at NUH. When we compare the Period 2 discharge scenario (Figure 17) with the *Israeli hospital* scenario, which uses a discharge distribution similar to the one at the Israeli hospital and keeps all other settings the same as in the baseline, we observe a significant improvement of waiting time statistics after early discharge. That is, the 6-hour service level is

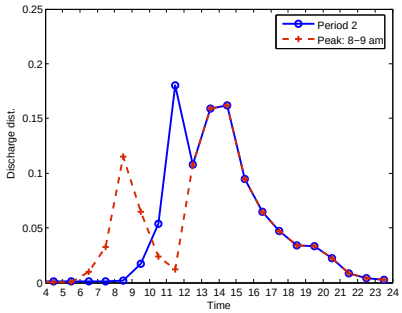
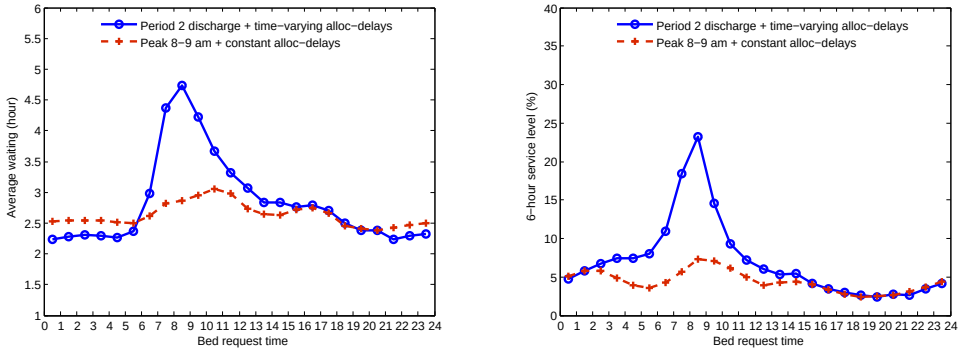


Figure 18 **Period 2 discharge distribution and a hypothetical Period 3 discharge distribution: first peak at 8-9am. In both distributions, 26% patients discharge before noon.**



(a) Average waiting time (b) 6-hour service level

Figure 19 **Period 2 policy: Period 2 discharge distribution and time-varying mean allocation delays; Period 3 policy: hypothetical discharge distribution with first peak at 8-9am and constant mean allocation delays.**

reduced from 9.29% in the Israeli hospital scenario to 5.50% in the Period 2 scenario (with the hourly peak value reduced from 45% to 23%); the average waiting time also reduces from 3.07 to 2.73 hours. Back to NUH, although our simulation suggests that Period 2 discharge policy has little impact on waiting time statistics for ED-GW patients, there could be other benefits that this paper has not modeled. For example, it is believed that the early discharge allows more flexibility to transfer patients from ICU to general wards when ICU wards become congested.

5.3. Period 3 policy has a significant impact on stabilizing waiting time statistics

Now we consider a hypothetical discharge distribution, which still discharges 26% of patients before noon as in Period 2, but shifts the first discharge peak time to 8-9am, i.e., three hours earlier than the first discharge peak time in Period 2. Figure 18 plots this hypothetical discharge distribution. In addition, we assume a hypothetical allocation delay model: each allocation delay (pre- or post-allocation delay) follows a log-normal distribution with a *constant* mean, which is estimated from the empirical daily average. The estimated means of the pre- and post-allocation delays are 1.07 and 1.20 hours, respectively. We keep the same values of CV as in the baseline scenario, i.e., CV = 1 and 0.6 for the pre- and post-allocation delays, respectively. We call the combination of the hypothetical discharge distribution and the hypothetical allocation delay model a *Period 3 policy*.

Under this Period 3 policy, the discharge distribution is more aggressive than the Period 2 one, and the allocation delays are stable with constant means throughout the day. For consistency, we call the combination of Period 2 discharge distribution and the time-varying allocation delay model (see Section 4.1) the *Period 2* policy. This is the policy that has been practiced at NUH since January 2010.

Figure 19 compares the hourly waiting time statistics for Period 2 and our hypothetical Period 3 policy. From the figure, we can see the hourly waiting time performances are almost stabilized throughout the day under the Period 3 policy. Patients requesting beds in the morning (7am to noon) now experience similar average waiting times as the daily average. The overall average waiting time drops from 2.73 hours under the Period 2 policy to 2.62 hours under the Period 3 policy, a 7-minute reduction. The 6-hour service level drops from 5.50% to 4.11%. The overall overflow proportion slightly drops, from 16.91% under the Period 2 policy to 16.34% under the Period 3 policy.

The Period 3 policy is purely hypothetical and may not be completely practical. Having the first discharge peak between 8 and 9am is challenging, since the morning rounds usually start at about this time, and it is known as the period of peak activities for ward nurses. Achieving stabilized allocation delays also needs coordination throughout the entire hospital and proper staffing at BMU, wards, and ED in various hour of the day. Though we recognize these difficulties, we believe that Period 3 policy provides a concrete direction that NUH and other hospitals can aim at to achieve time-stable waiting time performances.

In addition, the following procedures can potentially help move the first discharge peak time earlier. First, identify a few best candidates who can realistically be discharged earlier next day [26], and in the morning rounds, physicians focuses on them first. Second, initiate a nurse-led ward rounds/discharges, a recently emerged concept [36, 43], which can facilitate the discharge process (e.g., medically stable patients can be discharged without further discussion with physicians on the day of discharge [36]). More importantly, hospitals need to create compatible incentives for patients and their families, physicians, nurses and staff to work together to facilitate the discharge process. We leave these explorations for future study.

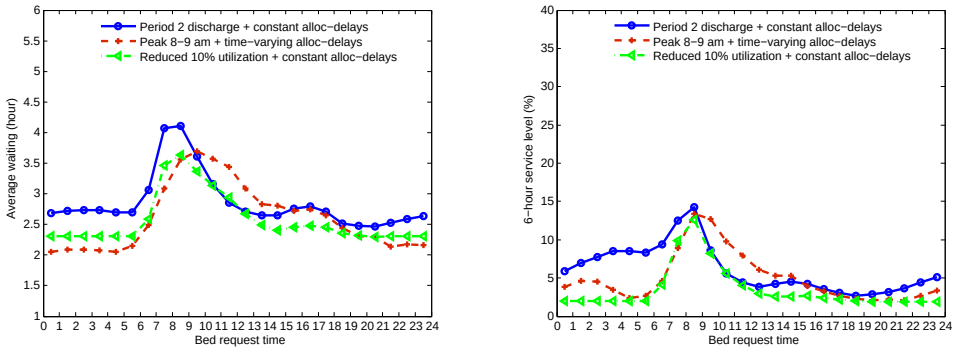
5.4. Scenarios which do not stabilize waiting time statistics

To achieve the stable performance in waiting time, the hypothetical Period 3 policy in Section 5.3 requires improvements in both the discharge time and allocation delay. In this section, we demonstrate that the simultaneous improvement is necessary. We also demonstrate that increasing bed capacity within a certain range, even when combined with a time-stable allocation delay, does not necessarily stabilize the waiting time statistics.

To show our results, we consider three scenarios. The first scenario uses Period 2 discharge distribution and assumes the constant-mean allocation delay model as in Period 3 policy. The second scenario uses the hypothetical Period 3 discharge distribution (red curve in Figure 18) and

assumes the time-varying allocation delay model. These two scenarios differ from the Period 3 scenario only in one factor: either the discharge distribution or the allocation delay model. In the third scenario, we use the baseline (Period 1) discharge distribution, assume the constant-mean allocation delay model, and increase the number of servers from 557 (baseline) to 626 so that the utilization rate is reduced from 90.2% to 80.3% (a 10% *absolute* reduction). All other settings not specified here remain the same as in the baseline for each of the three scenarios.

Figure 20 plots the waiting time statistics for these three scenarios. We see that in all three scenarios, the average waiting time is not stabilized, i.e., the average waiting time is still about 1-2 hours longer than the daily average for patients requesting beds between 7am and 11am. The hourly 6-hour service level, though, appears to be more stable than the average waiting time for each scenario, especially considering the peak value is 31% in the baseline.



(a) Average waiting time (b) 6-hour service level

Figure 20 **Hourly waiting time statistics under three scenarios. Scenario 1: Period 2 discharge distribution and constant mean allocation delays; Scenario 2: Period 3 discharge distribution and time-varying mean allocation delays; Scenario 3: 10% absolute reduction in utilization and constant mean allocation delays.**

Although the capacity increase in the third scenario does not lead to time-stable waiting time statistics, it results in a significant reduction in both the *daily* waiting time statistics and the overall overflow proportion. Under this scenario, we observe that (i) the average waiting time reduces from 2.81 hours (baseline) to 2.50 hours; (ii) the 6-hour service level reduces from 6.23% (baseline) to 2.87%; and (iii) the overflow proportion, dropping from 17.04% (baseline) to 8.36%, shows a 8.68% *absolute* reduction which is close to the 10% absolute reduction in utilization. The phenomenon that reducing utilization leads to a similar amount of absolute reduction in the overflow proportion actually holds in our simulation experiments with absolute utilization reduction ranging from 2% (the level achieved by NUH from 2008 to 2010) to 10%. Note that observation (iii) contrasts with the previous observation in Sections 5.2 and 5.3 that early discharge is of little use to reduce the overall overflow proportion, even under the hypothetical Period 3 policy. The intuitive explanation why early discharge policy alone fails to reduce overflow proportion is given in Section 11.4 of the online supplement [40].

To explore more scenarios in which the waiting time performances can or cannot be stabilized, we test other hypothetical discharge distributions combined with the time-varying or constant-mean allocation delay models. See Section 11.2 of the online supplement [40] for the experimented policies and simulation results. We highlight one finding here. If the hospital retains the first discharge peak time between 11am and noon in Period 2 while pushing for a 75% discharge before noon, the waiting time statistics cannot be stabilized even under the constant-mean allocation delay assumption. Comparing to Period 3 policy, this finding suggests that the timing of the first discharge peak has a larger impact than the proportion of discharge before noon on stabilizing waiting time performances.

6. Concluding remarks and future research

We have proposed a high-fidelity stochastic network model for inpatient flow management, which can be used as a tool to quantify the impact of various operational policies. In particular, the model captures the time-dependent waiting time performances for ED-GW patients and enables us to identify policies that can flatten the waiting times. Our model predicts that a hypothetical Period 3 policy can achieve time-stable waiting time statistics throughout the day, while it reduces the overflow proportion only slightly. Though we recognize the challenges in implementing the Period 3 policy in practice, we believe it can serve as a goal for hospital managers to aim at. Our model also predicts that increasing bed capacity, up to the level that utilization has a 10% absolute reduction, does not necessarily stabilize waiting time statistics, but can reduce the overflow proportion by around 9% (absolute amount).

Our model allows hospital managers to evaluate the trade-off between the benefit of reducing ED overcrowding and the cost from implementing a number of operational and strategic polices, including those that are not discussed in previous sections. For example, if a step-down care facility is established or expanded so that part of the social stays at NUH (defined as patients who stay more than 100 days in the inpatient wards for non-medical reasons) are eliminated, what will be the impact on the waiting time for admissions to wards? If the average LOS of PM-admitted patients can be reduced after a better coordination with other departments (e.g., radiology provides service after regular hours so the day of admission for some PM patients is no longer wasted), what will be the impact on the overall hospital performances? Our model can provide useful insights to decision makers in answering these kinds of questions.

Our model can potentially be extended in several directions. First, we use pre- and post-allocation delays as two black boxes to model all possible secondary bottlenecks including ward nurse and BMU staff shortage in certain time of a day. Detailed studies are needed to understand these secondary bottlenecks so that they can be explicitly incorporated into the model. The two-queue model proposed in [51] appears relevant to this line of research.

Second, currently ICU-type wards are not explicitly modeled, and therefore the congestion within ICU-type wards is not captured in our model. An extension is to model both ICU-type wards

and general wards as a stochastic network that has internal routings between these wards. The extended model can study the waiting times for ED-GW and ICU-GW patients as well as waiting times for ED-ICU patients or GW-ICU patients.

Third, considering the day-of-week phenomena is another important extension to make the model more realistic. Currently, we assume that EL patients have a stable daily arrival volume for each day of the week. Recent work pointed out that elective admission is the main source of daily occupancy variation in many hospitals [23]. Our model can be extended to predict the day-of-week performances and help design a better elective schedule.

Finally, there is a need to develop analytical methodology, not purely simulations, to predict performance measures that depend on hour-of-day.

References

[1] M. Armony, S. Israelit, A. Mandelbaum, Y. Marmor, Y. Tseytlin, and G. Yom-Tov, "Patient flow in hospitals: A data-based queueing perspective," 2011, working paper. [Online]. Available: http://www.stern.nyu.edu/om/faculty/armony/Patient%20flow%20main.pdf

[2] A. Bair, W. Song, Y. Chen, and B. Morris, "The impact of inpatient boarding on ed efficiency:a discrete-event simulation study," *J Med Syst*, vol. 34, pp. 919–929, 2010.

[3] S. L. Bernstein, D. Aronsky, R. Duseja, S. Epstein, D. Handel, U. Hwang, M. McCarthy, K. John McConnell, J. M. Pines, N. Rathlev, R. Schafermeyer, F. Zwemer, M. Schull, B. R. Asplin, and E. D. C. T. F. Society for Academic Emergency Medicine, "The effect of emergency department crowding on clinically oriented outcomes," *Academic Emergency Medicine*, vol. 16, no. 1, pp. 1–10, 2009.

[4] A. Birjandi and L. M. Bragg, *Discharge Planning Handbook for Healthcare: Top 10 Secrets to Unlocking a New Revenue Pipeline*. New York: Productivity Press, 2008.

[5] L. Brown, N. Gans, A. Mandelbaum, A. Sakov, H. Shen, S. Zeltyn, and L. Zhao, "Statistical analysis of a telephone call center," *Journal of the American Statistical Association*, vol. 100, no. 469, pp. 36–50, 2005.

[6] A. M. d. Bruin, A. v. Rossum, M. C. Visser, and G. Koole, "Modeling the emergency cardiac in-patient flow: An application of queuing theory," *Health Care Management Science*, pp. 1–13, 2006.

[7] Centers for Disease Control and Prevention, USA, "Health, United States," 2010. [Online]. Available: http://www.cdc.gov/nchs/data/hus/hus10.pdf

[8] C. Chan, G. Yom-Tov, and G. J. Escobar, "When to use Speedup: An Examination of Intensive Care Units with Readmissions," 2011, working paper. [Online]. Available: http://www.columbia.edu/~cc3179/ICU_fluid.pdf

[9] J. Cochran and A. Bharti, "Stochastic bed balancing of an obstetrics hospital," *Health Care Management Science*, vol. 9, no. 1, pp. 31–45, 2006.

[10] J. G. Dai and W. Lin, "Maximum pressure policies in stochastic processing networks," *Operations Research*, vol. 53, pp. 197–218, 2005.

[11] Department of Health, United Kingdom, "Achieving timely simple discharge from hospital: A toolkit for the multi-disciplinary team," 2004. [Online]. Available: http://www.dh.gov.uk/en/Publicationsandstatistics/Publications/PublicationsPolicyAndGuidance/DH_4088366

[12] D. Eitel and D. A. Samuelson, "OR in the ER," *OR/MS Today*, vol. 38, no. 4, August 2011. [Online]. Available: http://www.informs.org/ORMS-Today/Public-Articles/August-Volume-38-Number-4/OR-in-the-ER

[13] Z. Feldman, A. Mandelbaum, W. A. Massey, and W. Whitt, "Staffing of time-varying queues to achieve time-stable performance," *Management Science*, vol. 54, no. 2, pp. 324–338, 2008.

[14] P. Glynn, "Perspective on traffic modeling," Slides presented at the 2012 Stochastic Network Conference, MIT, June 2012.

[15] L. Green, "Queueing analysis in healthcare," in *Patient Flow: Reducing Delay in Healthcare Delivery*, ser. International Series in Operations Research and Management Science, R. W. Hall, Ed. Springer US, 2006, vol. 91, pp. 281–307.

[16] ——, 2012, private communications.

[17] ——, "How many hospital beds?" *Inquiry*, vol. 39, no. 4, pp. 400–412, 2002.

[18] J. Griffin, S. Xia, S. Peng, and P. Keskinocak, "Improving patient flow in an obstetric unit," *Health Care Manag Sci*, 2011.

[19] D. Gross and C. M. Harris, *Fundamentals of Queueing Theory*. New York: Wiley, 1985.

[20] M. J. Hall, C. J. DeFrances, S. N. Williams, A. Golosinskiy, and A. Schwartzman, "National hospital discharge survey: 2007 summary," *Natl Health Stat Report*, no. 29, pp. 1–20, 24, 2010.

[21] R. Hall, D. Belson, P. Murali, and M. Dessouky, "Modeling patient flows through the healthcare system," in *Patient Flow: Reducing Delay in Healthcare Delivery*, R. Hall, Ed. Springer, 2006.

[22] J. M. Harrison, "Brownian models of open processing networks: canonical representation of workload," *Annals of Applied Probability*, vol. 10, pp. 75–103, 2000, correction: **13**, 390–393 (2003).

[23] J. Helm and M. Van Oyen, "Design and optimization methods for elective hospital admissions," 2012, working paper.

[24] J. E. Helm, S. AhmadBeygi, and M. P. Van Oyen, "Design and analysis of hospital admission control for operational effectiveness," *Production and Operations Management*, vol. 20, no. 3, pp. 359–374, 2011.

[25] N. Hoot and D. Aronsky, "Systematic review of emergency department crowding: Causes, effects, and solutions." *Ann Emerg Med*, vol. 52, pp. 126–36, 2008.

[26] Hospitalist Management Advisor, "To free beds for new admissions, triage best candidates for early discharge," 2006. [Online]. Available: http://www.hcpro.com/content/62360.pdf

[27] E. Howell, E. Bessman, S. Kravet, K. Kolodner, R. Marshall, and S. Wright, "Active bed management by hospitalists and emergency department throughput." *Annals of Internal Medicine*, vol. 149, no. 11, pp. 804–810, 2008.

[28] S. H. Jacobson, S. N. Hall, and J. R. Swisher, "Discrete-event simulation of health care systems," in *Patient Flow: Reducing Delay in Healthcare Delivery*, ser. International Series in Operations Research and Management Science, R. W. Hall, Ed. Springer US, 2006, vol. 91, pp. 211–252.

[29] S. Khanna, J. Boyle, N. Good, and J. Lind, "Impact of admission and discharge peak times on hospital overcrowding," *Health Informatics: The Transformative Power of Innovation*, pp. 82–88, 2011.

[30] S.-H. Kim, C. Chan, M. Olivares, and G. J. Escobar, "ICU admission control: An empirical study of capacity allocation and its implication on patient outcomes," 2012, working paper. [Online]. Available: http://www.columbia.edu/~cc3179/ICUadm_2012.pdf

[31] E. Litvak, M. C. Long, A. B. Cooper, and M. L. McManus, "Emergency department diversion: Causes and solutions," *Academic Emergency Medicine*, vol. 8, no. 11, pp. 1108–110, 2001.

[32] Y. Liu and W. Whitt, "Stabilizing customer abandonment in many-server queues with time-varying arrivals," 2012, working paper. [Online]. Available: http://www.columbia.edu/~ww2040/LiuWhittStabilizing032812.pdf

[33] A. Mandelbaum, P. Momcilovic, and Y. Tseytlin, "On fair routing from emergency departments to hospital wards: QED queues with heterogeneous servers," *Management Science*, 2012.

[34] National University Hospital, "BMU training guide: Inpatient operations," December 2011.

[35] J. M. Pines, S. Iyer, M. Disbot, J. E. Hollander, F. S. Shofer, and E. M. Datner, "The effect of emergency department crowding on patient satisfaction for admitted patients," *Academic Emergency Medicine*, vol. 15, no. 9, pp. 825–831, 2008.

[36] A. Pottle, C. Hayes, S. Plater, and C. Ilsley, "Evaluation of the first nine months of a nurse-led ward-round for patients post percutaneous coronary intervention (pci) in a tertiary centre," August 2011. [Online]. Available: http://spo.escardio.org/AbstractDetails.aspx?id=98190&eevtid=48

[37] E. S. Powell, R. K. Khare, A. K. Venkatesh, B. D. Van Roo, J. G. Adams, and G. Reinhardt, "The relationship between inpatient discharge timing and emergency department boarding," *The Journal of Emergency Medicine*, 2011.

[38] S. Schneider, F. Zwemer, A. Doniger, R. Dick, T. Czapranski, and E. Davis, "Rochester, New York: a decade of emergency department overcrowding," *Academic Emergency Medicine*, vol. 8, no. 11, pp. 1044–1050, 2001.

[39] R. Serfozo, *Basics of applied stochastic processes*, ser. Probability and its Applications (New York). Berlin: Springer-Verlag, 2009.

[40] P. Shi, J. G. Dai, D. Ding, J. Ang, M. Chou, J. Xin, and J. Sim, "Online Supplement for "Hospital Inpatient Operations: Mathematical Models and Managerial Insights"," 2012.

[41] Sinagpore Ministry of Health, "Waiting time for admission to ward," May 16 2012. [Online]. Available: http://www.moh.gov.sg/content/moh_web/home/statistics/healthcare_institutionstatistics/Waiting_Time_for_Admission_to_Ward.html

[42] K. Teow, E. El-Darzi, C. Foo, X. Jin, and J. Sim, "Intelligent analysis of acute bed overflow in a tertiary hospital in singapore," *Journal of Medical Systems*, pp. 1–10, Janurary 2011.

[43] The Department of Health, Social Services and Public Safety, United Kindom, "Nurse led discharge and in reach," April 2006. [Online]. Available: www.dhsspsni.gov.uk/nurse_led_discharge_and_in_reach.pdf

[44] The Straits Times, "Shortage of hospital beds, so some ops delayed," Feb 2012. [Online]. Available: http://www.straitstimes.com/BreakingNews/Singapore/Story/STIStory_767553.html

[45] S. Thompson, M. Nunez, R. Garfinkel, and M. Dean, "Efficient short-term allocation and reallocation of patients to floors of a hospital during demand surges," *Operations Research*, vol. 57, no. 2, pp. 261 – 273, 2009.

[46] United States General Accounting Office, *Hospital emergency departments: crowded conditions vary among hospitals and communities*. Washington, D.C.: United States General Accounting Office, 2003.

[47] F. d. Vericourt and O. B. Jennings, "Nurse staffing in medical units: A queueing perspective," *Operations Research*, vol. 59, no. 6, pp. 1320–1331, 2011.

[48] M. J. Vermeulen, J. G. Ray, C. Bell, B. Cayen, T. A. Stukel, and M. J. Schull, "Disequilibrium between admitted and discharged hospitalized patients affects emergency department length of stay," *Annals of emergency medicine*, vol. 54, no. 6, pp. 794–804, 2009.

[49] H. J. Wong, D. Morra, M. Caesar, M. W. Carter, and H. Abrams, "Understanding hospital and emergency department congestion: An examination of inpatient admission trends and bed resources," *Canadian Journal of Emergency Medicine*, vol. 34, no. 1, pp. 18–26, 2010.

[50] D. A. Yancer, D. Foshee, H. Cole, R. Beauchamp, W. de la Pena, T. Keefe, W. Smith, K. Zimmerman, M. Lavine, and B. Toops, "Managing capacity to reduce emergency department overcrowding and ambulance diversions," *Jt Comm J Qual Patient Saf*, vol. 32, no. 5, pp. 239–45, 2006.

[51] N. Yankovic and L. V. Green, "Identifying good nursing levels: A queuing approach," *Operations Research*, vol. 59, no. 4, pp. 942–955, 2011.

[52] S. Zeltyn, Y. N. Marmor, A. Mandelbaum, B. Carmeli, O. Greenshpan, Y. Mesika, S. Wasserkrug, P. Vortman, A. Shtub, T. Lauterman, D. Schwartz, K. Moskovitch, S. Tzafrir, and F. Basis, "Simulation-based models of emergency departments: Operational, tactical, and strategic staffing," *ACM Trans. Model. Comput. Simul.*, vol. 21, no. 4, pp. 24:1–24:25, Sep. 2011.